

# Speed versus accuracy in visual search: Optimal performance and neural architecture

Bo Chen

Computation and Neural Systems,  
California Institute of Technology, Pasadena, CA, USA



Pietro Perona

Computation and Neural Systems,  
California Institute of Technology, Pasadena, CA, USA



Searching for objects among clutter is a key ability of the visual system. Speed and accuracy are the crucial performance criteria. How can the brain trade off these competing quantities for optimal performance in different tasks? Can a network of spiking neurons carry out such computations, and what is its architecture? We propose a new model that takes input from V1-type orientation-selective spiking neurons and detects a target in the shortest time that is compatible with a given acceptable error rate. Subject to the assumption that the output of the primary visual cortex comprises Poisson neurons with known properties, our model is an ideal observer. The model has only five free parameters: the signal-to-noise ratio in a hypercolumn, the costs of false-alarm and false-reject errors versus the cost of time, and two parameters accounting for nonperceptual delays. Our model postulates two gain-control mechanisms—one local to hypercolumns and one global to the visual field—to handle variable scene complexity. Error rate and response time predictions match psychophysics data as we vary stimulus discriminability, scene complexity, and the uncertainty associated with each of these quantities. A five-layer spiking network closely approximates the optimal model, suggesting that known cortical mechanisms are sufficient for implementing visual search efficiently.

## Introduction

One of the most useful functions of the visual system is searching for things, such as food, mates, and threats. This is a difficult task: The relevant objects, whose appearance may not be entirely known in advance, are often embedded in irrelevant clutter, whose appearance and complexity may also be unknown. Furthermore, time is of the essence: The ability to quickly detect objects of interest is an evolutionary advantage. Speed

comes at the cost of making mistakes. Shorter decision times imply collecting less signal and expose the animal to detection errors. Thus, it is critical that each piece of sensory information is used efficiently to produce a decision in the shortest amount of time while keeping the probability of errors within an acceptable limit.

Psychologists have characterized human visual search performance (Treisman & Gelade, 1980; Duncan & Humphreys, 1989; J. Palmer, 1994; Verghese & Nakayama, 1994; Carrasco & Yeshurun, 1998; Cameron, Tai, Eckstein, & Carrasco, 2004; Wolfe, Horowitz, & Kenner, 2005; Navalpakkam, Koch, Rangel, & Perona, 2010; Wolfe, Palmer, & Horowitz, 2010; Eckstein, 2011; Pomplun, Garaas, & Carrasco, 2013), quantified as response time (RT) and error rate (ER), in relation to properties of the search environment such as the distinctiveness of the target against the background clutter (Duncan & Humphreys, 1989; Verghese & Nakayama, 1994), the complexity of the image (J. Palmer, 1994; Carrasco & Yeshurun, 1998), and the likelihood that an object of interest may be present (Wolfe et al., 2005, 2010). As shorter RT implies higher ER, in order to achieve the best performance humans must trade off RT and ER. However, it is unknown what the optimal RT versus ER tradeoff should be in a given environment. It is also unknown whether human visual search performance is optimal.

Models of visual search fall into two categories. Stochastic accumulators were introduced to model discrimination (Ratcliff, 1985; Busemeyer & Townsend, 1993; Shadlen & Newsome, 2001; Usher & McClelland, 2001; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Brown & Heathcote, 2008) and visual search (Wolfe, 2007; Purcell, Schall, Logan, & Palmeri, 2012). The decision signal is either obtained from electrophysiological recordings from decision-implicated areas (e.g., frontal eye field: Woodman, Kang, Thompson, &

Citation: Chen, B., & Perona, P. (2015). Speed versus accuracy in visual search: Optimal performance and neural architecture. *Journal of Vision*, 15(16):9, 1–29, doi:10.1167/15.16.9.

doi: 10.1167/15.16.9

Received January 30, 2015; published December 16, 2015

ISSN 1534-7362 © 2015 ARVO

Schall, 2008; Heitz & Schall, 2012; Purcell et al., 2012; lateral intraparietal area: Mazurek, Roitman, Ditterich, & Shadlen, 2003; Wong, Huk, Shadlen, & Wang, 2007) or the result of an educated guess to fit the phenomenology (J. Palmer, Huk, & Shadlen, 2005; Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012). Stochastic accumulator models are appealing because of their conceptual simplicity and because they fit behavioral data well. However, these models do not attempt to explain search performance in terms of the underlying primary signals and neural computations.

Ideal observer models have been developed to study which computations and mechanisms may be optimal for visual discrimination (Geisler, 1989; J. Palmer et al., 2005) and visual search under fixed time presentations (J. Palmer, Verghese, & Pavel, 2000; Verghese, 2001; Geisler, 2011; Ma, Navalpakkam, Beck, Van Den Berg, & Pouget, 2011; Shimozaki, Schoonveld, & Eckstein, 2012) using signal detection theory (Green & Swets, 1966). This line of work leads us to the question of whether it is possible to derive an ideal observer for visual search that may predict simultaneously both RT and ER.

We propose a principled model for visual search and use it for studying the ER versus RT tradeoff that the brain faces. We take the Bayesian point of view: We model a system that through experience (or through evolution) is familiar with the statistics of the scene. The input to our system is an array of idealized cortical hypercolumns that, in response to a visual stimulus, produce firing patterns that are Poisson and conditionally independent. After this assumption is made the model that characterizes the optimal ER versus RT tradeoff is derived with no additional assumptions and no additional free parameters. Our model computes the optimal tradeoff given the firing patterns of V1 neurons and a probabilistic description of the task. Last, we are interested in understanding whether such an observer might plausibly be implemented by neural mechanisms such as a network of spiking neurons. We develop such an architecture and compare its performance with that of the optimal model and humans.

## Results

### Optimal ER versus RT tradeoff in visual search

#### Visual search setup

The goal of the observer is to detect the presence of a target object in a cluttered image as quickly and accurately as possible while maintaining fixation at the center of the image. The observer makes a binary decision between two categories of stimuli: target present ( $C = 1$ ) and target absent ( $C = 0$ ). When the

target is present, its location is not known in advance; it may be one of  $L$  locations in the image. The observer reports whether the target appears but not where.

In our experiments the target and distractor objects appear at  $M$  locations ( $M \leq L$ ) in each image, where  $M$  reflects the complexity of the image and is known as the *set size* (see Figure 1a). The objects are simplified to be oriented bars, and the only feature in which the target and distractor differ is orientation. Target distinctiveness is controlled by the difference in orientation between target and distractors, referred to as the *orientation difference* ( $\Delta\theta$ ). Prior to image presentation, the set of possible orientations for the target and the distractor is known, whereas the set size and orientation difference may be unknown and may change from one image to the next (see the Psychophysics section for details).

#### Sensory input

The optimal ER versus RT tradeoff is defined with respect to a specific set of assumptions regarding the sensory input. Below we outline these assumptions. Most visual search models assume that the input is a high-level decision-related signal (e.g., a Gaussian random walk with a category-dependent slope; J. Palmer et al., 2000; Verghese, 2001; Wolfe, 2007; Purcell et al., 2012) or the ER or RT statistics of a related visual task (e.g., visual discrimination; Doshier, Han, & Lu, 2004, 2010). By contrast, we prefer to analyze the more general setting and consider sensory input from the early stages of the visual system (retina, lateral geniculate nucleus, and primary visual cortex).

We choose to model the primary visual cortex as a collection of hypercolumns that comprise neurons producing action potentials that are distributed according to Poisson distributions. The firing rate of the neurons depends on the nature of the visual stimulus; conditioned on the stimulus, the neurons are independent and their firing rate does not decay in time. This is an instance of the linear–nonlinear Poisson (LNP) model (Chichilnisky, 2001; Simoncelli, Paninski, Pillow, & Schwartz, 2004), which is commonly used to model neural responses. The anatomy and physiology of these stages of the visual system are well characterized (Hubel & Wiesel, 1962). These mechanisms compute local properties of the image (e.g., color contrast, orientation, spatial frequency, stereoscopic disparity, and motion flow; Felleman & Van Essen, 1991) and communicate these properties to downstream neurons for further processing. Accordingly, we assume that the observer's decision is based on the sequence of action potentials from orientation-selective neurons in V1 (Hubel & Wiesel, 1962). The firing patterns of the neurons are modeled with a Poisson process (Sanger, 1996). While Gaussian firing rate

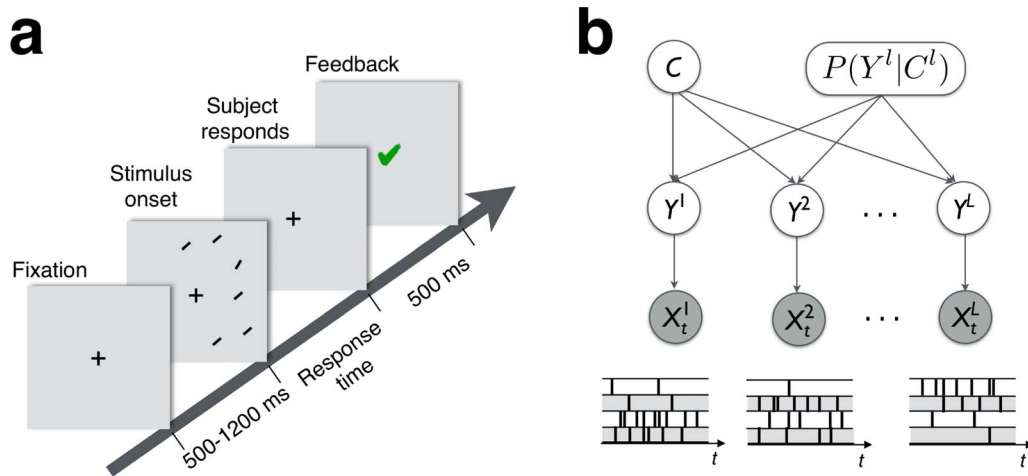


Figure 1. Visual search setup. (a) Each trial starts with a fixation screen. Next, the stimulus is displayed. The stimulus is an image containing  $M$  oriented bars that are positioned in  $M$  out of  $L$  possible display locations ( $M = 6$ ,  $L = 12$  in this example). One of the bars may be the target. The stimulus disappears as soon as the subject responds by pressing one of two keys to indicate whether a target was detected or not. Feedback on whether the response was correct is then presented on the screen, which concludes the trial. The subjects were instructed to maintain center fixation at all times and to respond as quickly and as accurately as possible. (b) A generative model of the stimulus. The stimulus class  $C$  and a prior distribution on the stimulus orientation  $P(Y^l|C^l)$  decide, for each display location  $l$ , the orientation  $Y^l$  (may be blank). The orientation  $Y^l$  determines in turn the observation  $X_t^l$ , which are firing patterns from a hypercolumn of V1 orientation-selective neurons at location  $l$  over the time window  $[0, t]$ . (The firing patterns of four neurons are shown at each location.)

models (Verghese, 2001) have also been used in the past, the Poisson model represents more faithfully the spiking nature of neurons (Sanger, 1996; Beck et al., 2008; Graf, Kohn, Jazayeri, & Movshon, 2011). In a Poisson process, the number  $n$  of events (i.e., action potentials) that will be observed during 1 s is distributed as  $P(n|\lambda) = \lambda^n e^{-\lambda} / n!$ , where  $\lambda$  is the expected number of events per second (e.g., the firing rate of the neuron).

The firing patterns are produced over the time interval  $[0, t]$  by a population of  $N$  neurons, also known as a *hypercolumn*, from each of the  $L$  display locations. Each neuron has a localized spatial receptive field and is tuned to local image properties (Hubel & Wiesel, 1962), which in our case is the local stimulus orientation; the preferred orientations of neurons within a hypercolumn are distributed uniformly in  $[0^\circ, 180^\circ)$ .  $\lambda_\theta^i$ , the expected firing rate of the  $i$ th neuron, is a function of the neuron's preferred orientation  $\theta_i$  and the stimulus orientation  $\theta \in [0^\circ, 180^\circ)$ :

$$\lambda_\theta^i = (\lambda_{\max} - \lambda_{\min}) \times \exp\left(-\frac{1}{2\psi^2} \left(\min_{k=-1,0,1} (|\theta - \theta_i + k180^\circ|)^2\right)\right) + \lambda_{\min} \quad (1)$$

(in spikes per second, or Hz), where  $\lambda_{\min}$  and  $\lambda_{\max}$  are a neuron's minimum and maximum firing rates, respectively, and  $\psi \in (0^\circ, 180^\circ)$  is the half tuning width.

Figure 4c shows the tuning functions of a hypercolumn

of eight neurons, and Figure 4f shows the sample spike trains from two locations with different local stimulus orientations.

Our analysis starts from simulations of the firing patterns of V1 hypercolumn according to the tuning curve described above. Analysis using electrophysiological recordings from V1 neurons (Graf et al., 2011) is left for future work.

### Optimal ER versus RT tradeoff

Under the assumptions regarding the sensory input described above, given the firing pattern of V1 neurons that is caused by the image, an observer faces a double decision. First, at each time instant it has to decide whether the information in the input collected so far is sufficient to detect the target. Second, once information is deemed sufficient, it has to decide whether the target is present or not. Since the neurons are noisy, the information collected over a finite-size interval is insufficient to achieve certainty, and any decision is subject to error. The longer the observer waits, the more information it collects and the fewer errors it will make. However, in many ecologically relevant situations, such as searching for food and avoiding predators, time is expensive. Thus, the observer must trade off the cost of making more errors with the cost of being slower. We clarify this concept in the following paragraphs.

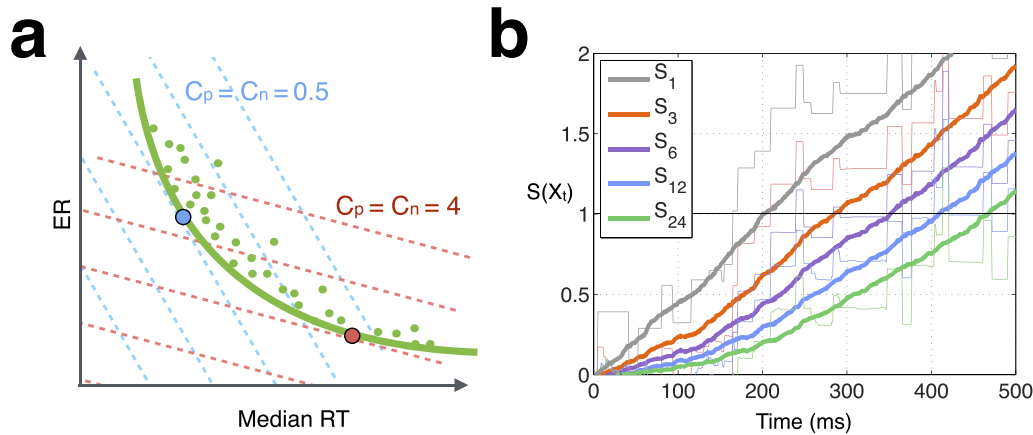


Figure 2. The optimal decision strategy for trading off ER versus RT. (a) The optimal decision strategy minimizes the Bayes risk (Equation 2). An observer's performance is indicated by a green dot in the ER versus median RT plane. The dashed lines indicate equicost contours for a given ratio of the errors versus time cost. Red lines correspond to relatively low error cost, while the blue lines correspond to high error cost (for simplicity the two error costs are equal here). The green curve indicates the performance of the observer using the optimal strategy (the lower envelope of all observers). The blue and red dots indicate the locus where the optimal curve is tangent to the blue and red equicost lines. Such points correspond to the setting of the observer that minimizes the Bayes risk. (b) The optimal strategy for visual search with set size  $M$  computes the log likelihood ratio  $S_M$  via a nonlinear combination of log likelihood ratios from all display locations (Equation 7). For difficult tasks such as  $M = 24$ ,  $S_M$  is clearly not a diffusion (e.g., initial slopes differ from later slopes), while  $S_1$  is exactly a diffusion because it corresponds to visual discrimination (Equation 4) rather than search. Unlike standard diffusion models, no parameter is needed to describe how  $S_M$  depends on  $M$ . Thick lines show averages over 200 target-present trials. Thin lines show the log likelihood ratios from a sample trial. (The same input pattern at the target location is used for different values of  $M$ .)

We seek the optimal decision strategy for trading off ER versus RT given the available information in the input. Optimality is measured in terms of the Bayes risk (Wald & Wolfowitz, 1948; Busmeyer & Rapoport, 1988):

$$\text{BayesRisk} = \mathbb{E}[\text{RT}] + C_p \mathbb{E}[\text{Declare } C = 1 | C = 0] + C_n \mathbb{E}[\text{Declare } C = 0 | C = 1], \quad (2)$$

where  $\mathbb{E}[\text{RT}]$  is the expected RT;  $\mathbb{E}[\text{Declare } C = 1 | C = 0]$  is the probability of the observer making a target-present decision when the target is absent, or the false-positive rate; and likewise,  $\mathbb{E}[\text{Declare } C = 0 | C = 1]$  is the false-negative rate.  $C_p$  and  $C_n$  are two free parameters: the cost (in seconds) of false-positive errors and the cost of false-negative errors, respectively. For example,  $C_p$  might be quantified in terms of the time wasted exploring an unproductive location while foraging for food, and  $C_n$  may be the time it takes to move to the next promising location. The relative cost of errors and time is determined by the circumstances in which the animal operates. For example, an animal searching for scarce food while competing with conspecifics will face a high cost of time (e.g., any delay in pecking a seed will mean that the seed is lost) and low cost of error (e.g., pecking on a pebble rather than a seed just means that the pebble can be spat out). Conversely, an airport luggage inspector faces high

false-reject error costs and comparatively lower time costs.  $C_p$  and  $C_n$  determine how often the observer is willing to make one type of error versus the other and versus waiting for more evidence. Thus, the Bayes risk measures the combined RT and ER costs of a given search mechanism. Given a set of inputs, the optimal strategy is the mechanism minimizing such cost (Figure 2a).

It may be shown (Chen & Perona, 2014) that although the optimal decision strategy for general visual search is often too expensive to compute exactly, in common settings its performance is indistinguishable from that of the Sequential Probability Ratio Test (SPRT) (Wald, 1945), which may be evaluated efficiently. Thus, while the SPRT is not strictly optimal for visual search (Chen & Perona, 2014), we still refer to it as the optimal strategy, bearing in mind that it is optimal for all practical purposes.

Call  $X_t$  the input observations, which is the collection of firing patterns of the V1 hypercolumn neurons from all display locations collected over the time window  $[0, t]$ . The SPRT takes the following form:

$$S(X_t) \triangleq \log \frac{P(C = 1 | X_t)}{P(C = 0 | X_t)} \begin{cases} \geq \tau_1 & \text{Declare target present} \\ \leq \tau_0 & \text{Declare target absent} \end{cases} \quad (3)$$

It considers  $S(X_t)$ , the log ratio of target-present ( $C = 1$ ) versus target-absent ( $C = 0$ ) probability given the



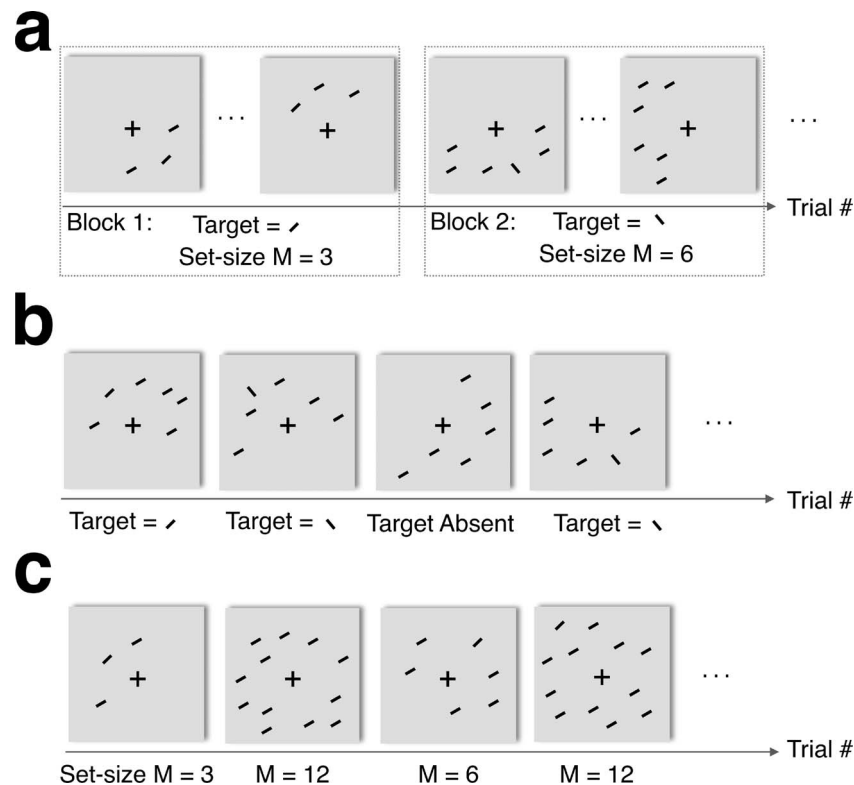


Figure 3. Common visual search settings. (a) Blocked. The orientation difference and the set size remain constant within a block of trials (outlined in dashed box) and vary only between blocks. (b) Mix contrast. The target orientation varies independently from trial to trial while the distractor orientation and the set size are held constant. (c) Mix set size. The set size is randomized between trials while the target and distractor orientations are fixed.

observations  $X_t$ . A target-present decision is made as soon as  $S(X_t)$  crosses an upper threshold  $\tau_1$ , while a target-absent decision is made as soon as  $S(X_t)$  crosses a lower threshold  $\tau_0$ . Until either event takes place, the observer waits for further information. For convenience we use base 10 for all our logarithms and exponentials; that is,  $\log(x) \triangleq \log_{10}(x)$  and  $\exp(x) \triangleq 10^x$ .

The thresholds  $\tau_1 > 0$  and  $\tau_0 < 0$  control the maximum tolerable ERs. For example, if  $\tau_1 = 2$  (i.e., a target-present decision is taken when the stimulus is  $>10^2$  times more likely to be a target than a distractor), then the maximum false-positive rate is 1%; Similarly If  $\tau_0 = -3$ , then target likelihood is  $<10^{-3}$  times the distractor's and the false-negative rate is at most 0.1%.  $\tau_1$  and  $\tau_0$  are judiciously chosen by the observer to minimize the Bayes risk in Equation 2 and hence are functions of the costs of errors. For example, if  $C_p > C_n$ , the observer should be less reluctant to make a false-negative error and thus should set  $|\tau_0| < \tau_1$ . In addition, if both  $C_p$  and  $C_n$  are large, the observer should increase  $|\tau_0|$  and  $\tau_1$  so that fewer errors are made in general at the price of a longer RT. Given this relationship, we parameterize the SPRT with the thresholds  $\tau_0$  and  $\tau_1$  instead of the costs of errors  $C_p$  and  $C_n$ .

Therefore, the optimal strategy for trading off ER and RT computes decisions using the SPRT (Wald,

1945), which compares the log likelihood ratio  $S(X_t)$  between target present and target absent with a pair of thresholds  $\tau_0$  and  $\tau_1$ . Next we explain how  $S(X_t)$  is computed.

### Computing the log likelihood ratio from sensory input

$S(X_t)$  can be systematically constructed from the visual input according to the graphical model in Figure 1b and can account for a wide variety of visual search tasks. We derive a general model that is capable of handling unknown set sizes and orientation differences in the independent and identically distributed (i.i.d.)-distractor heterogeneous search and Heterogeneous search sections. (Readers interested only in this general model are encouraged to skip to those sections.) To build up the concept, we start by reviewing models for simpler tasks including visual discrimination and visual search with known set sizes and orientation differences, both of which have already been explored in the literature (Wald, 1945; Chen, Navalpakkam, & Perona, 2011; Ma et al., 2011). We review that for simple discrimination,  $S(X_t)$  is a simple diffusion, and the optimal strategy is a diffuse-to-bound system (Ratcliff, 1985). We also show that in all other scenarios,  $S(X_t)$  is a nonlinear function of the input, and thus diffusions

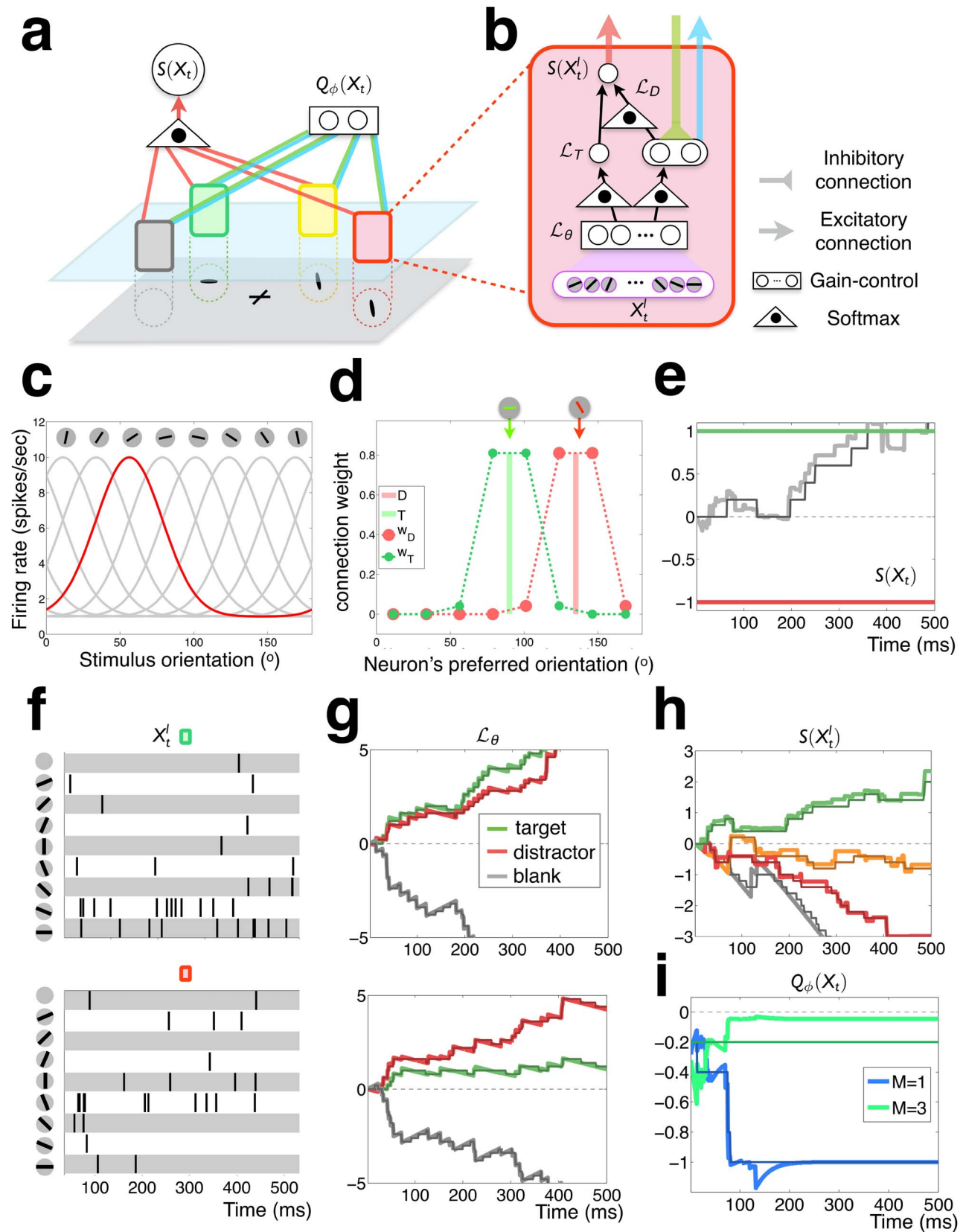


Figure 4. SPRT for heterogeneous visual search and its spiking network implementation. (a) SPRT for heterogeneous visual search is implemented by a five-layer network. It has two global circuits: One computes the global log likelihood ratio  $S(X_t)$  (Equation 10) from local circuits that compute log likelihood ratios  $\{S(X_t^i)\}_i$  (Equation 11), and the other estimates scene complexity  $Q_\phi(X_t)$  (Equation 25)

→

←

via gain control.  $Q_\phi(X_t)$  feeds back to the local circuit at each location. (b) The local circuit that computes the log likelihood ratio  $S(X_t^l)$ . Spike trains  $X_t$  from V1/V2 orientation-selective neurons are converted to log likelihood for task-relevant orientations  $\mathcal{L}_\theta$  (Equation 12). The log likelihoods of the distractor  $\mathcal{L}_D$  (second line of Equation 9) under every putative CDD are compiled together, sent (blue outgoing arrow) to the global circuit, and inhibited (green incoming arrow) by the CDD estimate  $Q_\phi$  (details in Equation 25). (c) Orientation tuning curves  $\lambda_\theta^l$  (Equation 1) of a hypercolumn consisting of  $N = 8$  neurons with half tuning width  $\psi = 22^\circ$ , minimum firing rate  $\lambda_{\min} = 1$  Hz, and maximum firing rate  $\lambda_{\max} = 10$  Hz. (d) Connection weights (Equation 18) between hypercolumn neurons and neurons coding for the distractor ( $W_D$ ) and the target ( $W_T$ ) for an observer searching for a  $90^\circ$  target (green bar) among  $135^\circ$  distractors (red bar; orientation difference =  $45^\circ$ ). (e–i) An instantiation of the signals propagating through the network in panel a with an orientation difference of  $45^\circ$  and two possible set sizes (1 and 3). (e) The log likelihood ratio  $S(X_t)$  computed using SPRT (black line) and the spiking implementation (gray line) reach the identical decision at similar RTs (350 ms). (f) Spike trains  $X_t^l$  at the target location (green box in panel a) and a distractor location (red box). (g) The corresponding orientation log likelihoods  $\mathcal{L}_\theta$ . (h) Local log likelihood ratios for the four color-coded locations in panel a. (i) Log likelihoods of the two CDDs—one for each putative set size. In panels g through i, darker lines on top of the lighter lines correspond to the spiking network approximation to SPRT.

are not optimal. All tasks considered are summarized in Table 1.

Let  $X_t^l$  denote the activity of the neurons at location  $l$  during the time interval  $[0, t]$  in response to a stimulus presented at time 0.  $X_t = \{X_t^l\}_{l=1}^L$  is the ensemble responses of all neurons from all locations. We have assumed that such responses are action potentials (Chen et al., 2011), but the same analysis also applies to analog response signals (e.g., a Gaussian random walk for each neuron; Verghese, 2001). Let  $\mathcal{L}_\theta(X_t^l) \triangleq \log P(X_t^l | Y^l = \theta)$  denote the log likelihood of the spike train data  $X_t^l$  when the object orientation  $Y^l$  at location  $l$  is  $\theta$ . Each  $\mathcal{L}_\theta(X_t^l)$  may be computed by a diffusion (Ratcliff, 1985) in which every new observation induces an additive update in  $\mathcal{L}_\theta(X_t^l)$  (for details, see the Spiking network implementation section; Equation 12).

**Homogeneous visual discrimination:** First consider the case where either the target or the distractor can appear at only one display location ( $L = M = 1$ ) and the target and distractor have distinct and unique orientations  $\theta_T$  and  $\theta_D$ , respectively. The visual system needs to determine whether the target or the distractor is present in the test image. The log likelihood ratio in this case is well known (Wald, 1945; rederived in Equation 19):

(Homogeneous Discrimination)

$$S(X_t) = \mathcal{L}_{\theta_T} - \mathcal{L}_{\theta_D}(X_t), \quad (4)$$

which, as first pointed out by Stone (1960), may be computed by a diffuse-to-bound mechanism (Ratcliff, 1985). In addition, as shown by Wald (1945), SPRT is optimal in minimizing the Bayes risk in Equation 2.

**Heterogeneous visual discrimination:** In a more general setting, both the target and the distractor could take one of multiple orientations. Use  $\Theta_T$  to denote the set of orientations that the target may take, and use  $\Theta_D$  to denote the set of orientations for the distractor. Further, use  $n_T$  and  $n_D$  to denote the number of orientations in set  $\Theta_T$  and  $\Theta_D$ , respectively. We call heterogeneous visual discrimination the case where  $n_T > 1$  and/or  $n_D > 1$ . The log likelihood ratio is (Ma et al., 2011; rederived in Equation 20)

(Heterogeneous Discrimination)

$$S(X_t) = \underset{\theta \in \Theta_T}{\text{Smax}} (\mathcal{L}_\theta(X_t) - \log(n_T)) - \underset{\theta \in \Theta_D}{\text{Smax}} (\mathcal{L}_\theta(X_t) - \log(n_D)), \quad (5)$$

where  $\text{Smax}(\cdot)$  is the softmax function. For a vector  $\mathbf{v}$  and a set of indices  $\mathcal{I}$ ,

Task	$L$	Set size, $M$	$\theta_T$ and $\theta_D$	CCD (distribution of distractor orientation)	$S(X_t)$ expression
Homogeneous discrimination	1	Known, $M = 1$	Known	Known	Equation 4
Heterogeneous discrimination	1	Known, $M = 1$	Unknown	Known	Equation 5
Homogeneous search	$>1$	Known, $M = L$	Known	Known	Equation 7
i.i.d.-distractor heterogeneous search	$>1$	Known, $M = L$	Unknown	Known	Equation 8
Heterogeneous search	$>1$	Unknown, $1 < M \leq L$	Unknown	Unknown	Equation 10

Table 1. Visual discrimination and visual search tasks discussed in this article.  $L$  = number of total display locations;  $M$  = number of display items;  $\theta_T$  and  $\theta_D$  = target and distractor orientations, respectively. We use *known* and *unknown* to refer to whether a quantity is known at stimulus onset. In many tasks,  $\theta_T$  and  $\theta_D$  are unknown but are sampled according to a distribution. Such distribution  $\phi$  is called a *conditional distractor distribution*, where  $\phi_\theta = P(Y^l = \theta | C^l = 0)$  for any location  $l$ .  $S(X_t) = \log P(C = 1 | X_t) / P(C = 0 | X_t)$  is the class log posterior ratio that SPRT computes. Our model accounts for the heterogeneous search task, which subsumes all other tasks on the list.

$$\text{Smax}(\mathbf{v}) \triangleq \log \sum_{i \in \mathcal{I}} \exp(v_i). \quad (6)$$

Softmax can be thought of as the marginalization operation in log probability space. It computes the log probability of a set of mutually exclusive events from the log probabilities of the individual events. For example, for two mutually exclusive events ( $A_1$ ,  $A_2$ ), we have  $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ , then  $\log P(A_1 \cup A_2) = \text{Smax}_{i=1,2}(\log P(A_i))$ . Since the different target orientations are mutually exclusive, their log likelihoods should be combined using the softmax function to compute the log likelihood for the target. The same argument applies to the distractor.

It is important to note that the log likelihood ratio for heterogeneous discrimination is not a diffusion. Rather, it combines diffusions in a nonlinear fashion (via a softmax). Diffuse to bound (Ratcliff, 1985) does not give the optimal decision mechanism here or in any of the settings we discuss later. Moreover, while a diffusion model may require additional parameters specifying how the statistics of the diffusions relate to the task parameters (set size in this case; J. Palmer et al., 2005; Drugowitsch et al., 2012), the construction of SPRT is parameter free, as shown in Figure 2b. In the Psychophysics section, we see that SPRT can generalize to novel experimental settings, which is nontrivial for diffusion models.

**Homogenous search:** Now that we have analyzed the case of discrimination (one item visible at any time) we explore the case of search (multiple items present simultaneously, one of which may be the target). Consider the case where all the  $L$  display locations are occupied by either a target or a distractor (i.e.,  $L = M > 1$ ) and the display contains either one target or none. The target orientation  $\theta_T$  and the distractor orientation  $\theta_D$  are again unique and known; that is,  $n_T = n_D = 1$ . The log likelihood ratio of target present versus target absent is given by Chen et al. (2011; rederived in Equation 21):

(Homogeneous Search)

$$S(X_t) = \text{Smax}_{l=1,\dots,L} \left( S(X_t^l) - \log(L) \right), \quad (7)$$

where  $S(X_t^l) = \mathcal{L}_{\theta_T}(X_t^l) - \mathcal{L}_{\theta_D}(X_t^l)$  is the log likelihood ratio for homogenous discrimination at location  $l$  (see Equation 4).  $S(X_t)$  combines the local log likelihood ratio  $S(X_t^l)$  from all locations using a softmax because the target can appear at only one of  $L$  disjoint locations.

**Independent and identically distributed (i.i.d.)-distractor heterogeneous search:** Now we describe our general model of visual search. We start with the simple case where the set size is known ( $M = L > 1$ ) but the orientation difference is not ( $n_T > 1$  and/or  $n_D > 1$ ). In addition, we assume that target and distractor orien-

tations are sampled i.i.d. in space according to some distribution. We refer to this as the *i.i.d.-distractor heterogeneous search*.

We call a conditional distractor distribution (CDD) the distribution of orientation  $Y^l$  at any nontarget location  $l$ ; that is,  $P(Y_l | C^l = 0)$ . We denote CDD with  $\phi$ , where  $\phi_\theta \triangleq P(Y_l = \theta | C^l = 0)$ . Thus,  $\phi$  is a  $n_D$ -dimensional probability vector; that is, each element of  $\phi$  is nonnegative and all elements sum to 1. We introduce CDD here because it is a key element in the general model of visual search, as becomes clear later. In contrast, the conditional target distribution  $P(Y_l = \theta | C^l = 1)$  is not as vital and is assumed to be uniform for notation clarity (see Equation 26 for cases with general target distributions and different CDDs over locations).

The log likelihood ratio may be computed as

(i.i.d. – Distractor Heterogeneous Search)

$$S(X_t) = \text{Smax}_{l=1,\dots,L} \left( S(X_t^l) - \log(L) \right), \quad (8)$$

$$\text{where } S(X_t^l) = \text{Smax}_{\theta \in \Theta_T} \left( \mathcal{L}_\theta(X_t^l) - \log(n_T) \right) - \text{Smax}_{\theta \in \Theta_D} \left( \mathcal{L}_\theta(X_t^l) + \log \phi_\theta \right). \quad (9)$$

The log likelihood ratio expressions (Equations 8 and 9) are obtained by nesting appropriately the models of homogeneous search and heterogeneous discrimination. At the highest level is the softmax over locations, as in Equation 7. At each location  $l$ ,  $S(X_t^l)$  is obtained as the difference between the log likelihood of the target and that of the distractor (Equation 9), which is reminiscent of Equation 5. Computing the target log likelihood requires marginalizing over the unknown target orientation with a softmax (again assuming uniform prior over possible target orientations in  $\Theta_T$ ). Similarly, the distractor log likelihood marginalizes over the distractor orientation according to the CDD. **Heterogeneous search:** Finally, in the most ecologically relevant situations the complexity and target distinctiveness are not known in advance. In other words, all search parameters  $M$ ,  $\theta_T$ , and  $\theta_D$  are stochastic ( $n_T$  and/or  $n_D > 1$ ). This scenario may be handled using mechanisms for i.i.d. distractor heterogeneous search above as building blocks. For example, for a fixed set size, each nontarget location has a certain probability of being blank (as opposed to containing a distractor), which is captured by the CDD. When set size changes, CDD will change correspondingly. Therefore, knowing the CDD effectively allows us to infer the set size and vice versa. Our strategy is to infer the CDD along with the class variables using Bayesian inference.

Let  $P(\phi)$  be the prior distribution over the CDDs  $\phi$ . Note that, technically,  $P(\phi)$  is a distribution over distributions. Computing the log likelihood ratio



requires marginalizing out  $\phi$  according to  $P(\phi)$  and the observation  $X_t$ . We assume that the observer has been exposed to this task for some time and has estimated  $P(\phi)$ . We also assume that the target distribution is independent of the CDD (and relax this assumption in Equation 29). The log likelihood ratio is (see derivations in Materials and method)

(General Model: Heterogeneous Search)

$$S(X_t) = \text{Smax}_{l=1\dots L} \left( S(X_t^l) - \log(L) \right), \quad (10)$$

where

$$\begin{aligned} S(X_t^l) = & \text{Smax}_{\theta \in \Theta_T} \left( \mathcal{L}_\theta(X_t^l) - \log(n_T) \right) \\ & + \text{Smax}_{\phi \in \Phi} \left( - \text{Smax}_{\theta \in \Theta_D} \left( \mathcal{L}_\theta(X_t^l) + \log \phi_\theta \right) \right. \\ & \left. + Q_\phi(X_t) \right), \end{aligned} \quad (11)$$

where  $Q_\phi(X_t) \triangleq \log P(\phi|X_t)$  is the log posterior of the CDDs given the observations  $X_t$  (see below). The only difference between Equations 10 and 11 and those describing the i.i.d.-distractor heterogeneous search (Equations 8 and 9) is the second line of Equation 11, where the CDD is marginalized out with respect to  $Q_\phi(X_t)$ . Since both the CDD  $\phi$  and the distractor orientation  $Y^l$  must be marginalized, two softmaxes are necessary. The equations do not explain how to compute  $Q_\phi(X_t)$ . It may be estimated simultaneously with the main computation by a scene complexity mechanism that is derived from first principles of Bayesian inference (see Equation 25). This mechanism extends across the visual field and may be interpreted as wide-field gain control (see Figure 4a and Equation 32).

A simpler alternative to inferring the CDD on a trial-by-trial basis is to ignore its variability completely by always using the same CDD obtained from the average complexity and target distinctiveness [i.e.,  $\mathbb{E}(\phi)$ ; see Equation 28]. This approach is suboptimal. Intuitively, if the visual scene switches randomly between being cluttered and sparse, then always treating the scene as if it had medium complexity would be either overly optimistic or overly pessimistic. Crucially, the predictions of this simple model are inconsistent with the behavior of human observers.

The general formulation in Equation 10 can describe the optimal strategy of a wide range of tasks such as visual search with unknown target appearance (Davis & Graham, 1981; Eckstein & Abbey, 2001; Hodsoll & Humphreys, 2001; see Figure 3b), unknown distractor appearance that is identical across locations (Duncan & Humphreys, 1989; Nagy, Neriani, & Young, 2005), unknown distractor appearance that is uniformly distributed everywhere (Duncan & Humphreys, 1989; Rosenholtz, 2001; Avraham, Yeshurun, & Linden-

baum, 2008; Chen & Perona, 2014), and unknown image complexity (Carrasco & Yeshurun, 1998; J. Palmer et al., 2000; Wolfe et al., 2010; see Figure 3c). Examples of describing each task above in our formulation are given in the Materials and method section.

**Search for multiple targets:** In Materials and method and Equation 31 we show that our visual search strategy may be generalized to detect the presence of multiple targets. While the calculations are different, the final expression is remarkably similar to Equations 10 and 11.

In conclusion, the optimal strategy for visual search—one that minimizes RTs while keeping ERs within an acceptable bound—is the SPRT conducted on the ratio  $S(X_t)$  defined in Equation 10. This strategy may be computed by nested combination of diffusions, as shown in Equations 10 and 11. In the next sections we explore the nature of  $X_t$  in the visual system and show that a simple network of spiking neurons may implement a close approximation of such a strategy.

## Spiking network implementation

### Local log likelihoods

We first explain how to compute  $\mathcal{L}_\theta(X_t)$ , the local log likelihood of the stimulus taking on orientation  $\theta$ , from spiking inputs  $X_t$  from V1.  $\mathcal{L}_\theta(X_t)$  is the building block of  $S(X_t)$  (Equation 4). Consider one spatial location, corresponding to a hypercolumn containing  $N$  neurons. Let  $K_t$  be the number of action potentials that were produced by all the neurons at that location up to time  $t$ ,  $\lambda_\theta^i$  denote the firing rate of neuron  $i$  when the stimulus orientation is  $Y = \theta$  (Figure 4d, f),  $t_s$  be the time at which the  $s$ th action potential takes place, and  $i(s)$  be the index of the neuron that produced it. Then the log likelihood of a set of action potentials  $X_t = \{t_s\}_{s=1}^{K_t}$  is (Jazayeri & Movshon, 2006; Chen et al., 2011; see Equation 18 for detailed derivations)

$$\mathcal{L}_\theta(X_t) \triangleq \log P(X_t|Y = \theta) = \sum_{s=1}^{K_t} \log \lambda_\theta^{i(s)} + \text{const.} \quad (12)$$

The first term is a diffusion, where each spike causes a jump in  $\mathcal{L}_\theta$ . This term can be implemented by integrate-and-fire (Dayan & Abbott, 2003) neurons—one for each relevant orientation  $\theta \in \Theta_T \cup \Theta_D$ —that receive afferent connections from all hypercolumn neurons with connection weights  $w_\theta^i = \log \lambda_\theta^i$  (Figure 4d). The second term is computationally irrelevant because it does not depend on the stimulus orientation  $\theta$  and it cancels with similar terms in Equation 11; it may be removed by a gain-

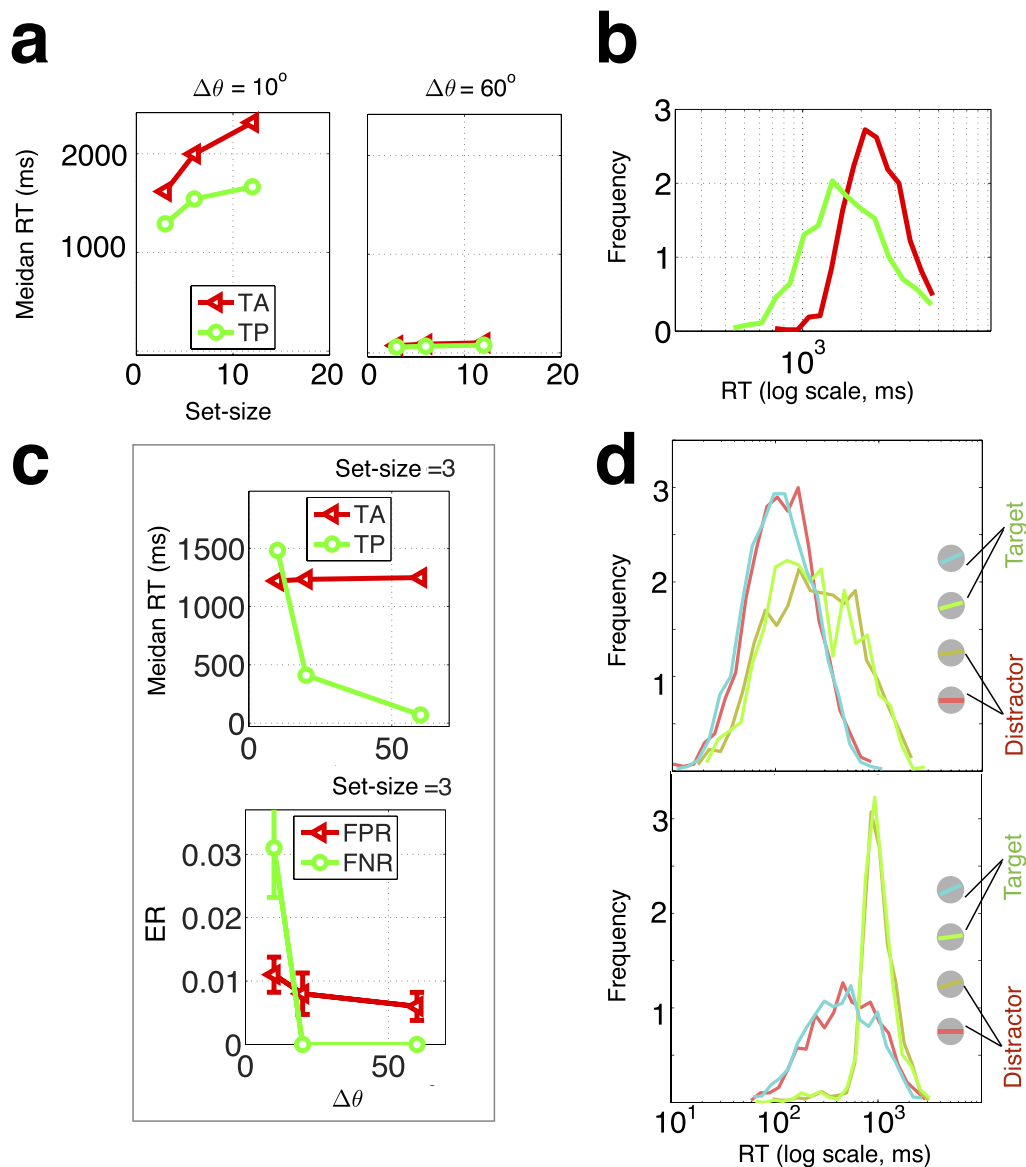


Figure 5. Qualitative predictions of SPRT (Simulations 1 through 3). (a) Set size effect on median RT under the blocked design (Simulation 1). SPRT predicts a linear RT increase with respect to set size when the orientation difference is low ( $10^\circ$ , left) and a constant RT when the orientation difference is high ( $60^\circ$ , right). The target-absent (TA) RT slope is roughly twice that of the target-present (TP) RT slope. (b) RT histogram under the blocked design with a  $10^\circ$  orientation difference and a set size of 12 items. RT distributions are approximately log normal. (c) Median RT (upper) and ER (lower) for visual search with heterogeneous target and distractor, mixed design (Simulation 2). (d) RT distributions of a visual discrimination task under different stimulus orientations (color coded) for two partitions of target and distractor orientations to achieve 1% overall error ( $\tau_1 = -\tau_0 = 2$ ; Simulation 3). In consecutive partition (upper),  $\Theta_0 = \{0^\circ, 12^\circ\}$ ,  $\Theta_1 = \{25^\circ, 37^\circ\}$ . In alternating partition (lower),  $\Theta_0 = \{0^\circ, 25^\circ\}$ ,  $\Theta_1 = \{12^\circ, 37^\circ\}$ . All simulations are generated with  $N = 16$  hypercolumn neurons, minimum and maximum firing rates  $\lambda_{\min} = 1$  Hz and  $\lambda_{\max} = 25$  Hz, half tuning width  $\psi = 45^\circ$ , and decision thresholds  $\tau_1 = -\tau_0 = 2$ .

control mechanism to prevent the dynamic range of membrane potential from exceeding its physiological limits (see Equation 32; Carandini, Heeger, & Movshon, 1999). Specifically, one may subtract from each  $\mathcal{L}_\theta$  a common quantity—for example, the average value of the all the  $\mathcal{L}_\theta$ s—without changing  $S(X_t^I)$  in Equation 11.

### Signal transduction

The log likelihood  $\mathcal{L}_\theta$  must be transmitted downstream for further processing. However,  $\mathcal{L}_\theta$  is a continuous quantity, whereas the majority of neurons in the central nervous system are believed to communicate via action potentials. We explored whether this communication may be implemented using action potentials

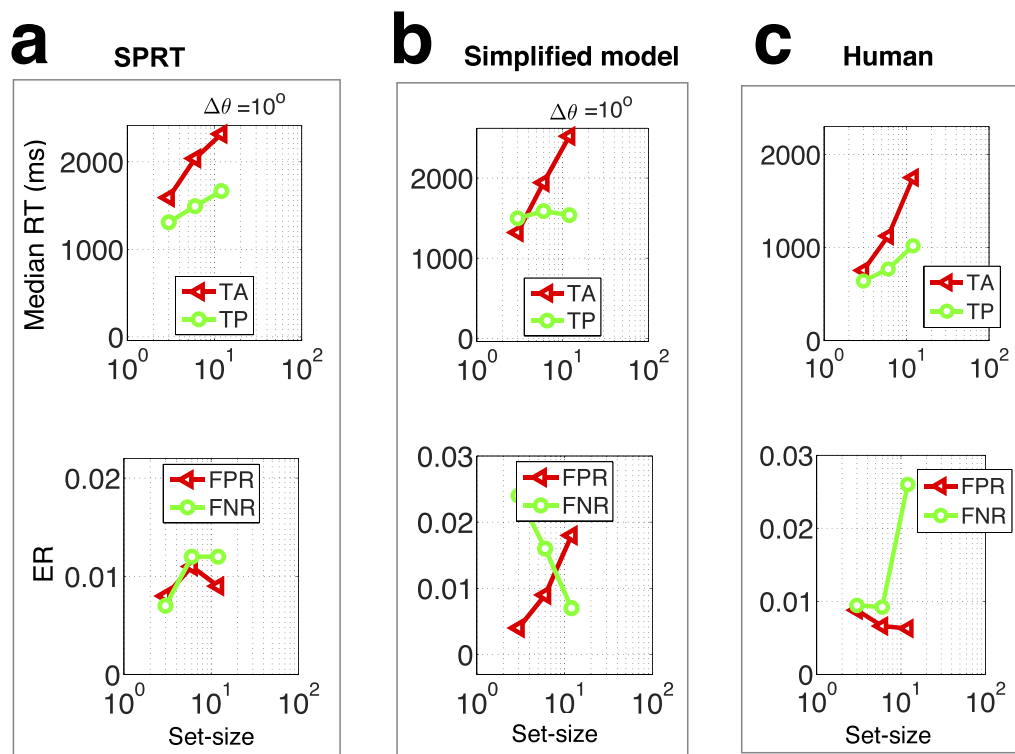


Figure 6. Qualitative model predictions and psychophysics data on visual search with unknown set size (Simulation 4). Median RT (upper) and ER (lower). False-positive rate (FPR) and false-negative rate (FNR) of visual search with homogenous target and distractor and unknown set sizes (Simulation 4) under two models: (a) SPRT that estimates the scene complexity parameter  $\phi$  (essentially the probability of a blank at any nontarget location) on a trial-by-trial basis (Equation 10) using a wide-field gain-control mechanism (Equation 25), and (b) a simplified observer that uses average scene complexity  $\hat{\phi}$  for all trials (Equation 28). Psychophysical measurements on human observers (Wolfe et al., 2010; spatial configuration search in figure 2-3, reproduced with permission here as panel c) are consistent with the optimal model (panel a). Simulation parameters are identical to those used in Figure 5.

(Gray & McCormick, 1996) emitted from an integrate-and-fire neuron. Consider a sender neuron communicating its membrane potential to a receiver neuron. The sender may emit an action potential whenever its membrane potential surpasses a threshold  $\tau_s$ . After firing, the membrane potential drops to its resting value and the sender enters a brief refractory period, the duration of which (about 1 ms) is assumed to be negligible. If the synaptic strength between the two neurons is also  $\tau_s$ , the receiver may decode the signal by simply integrating such weighted action potentials over time. This coding scheme loses some information due to discretization. Varying the discretization threshold  $\tau_s$  trades off the quality of transmission with the number of action potentials; a lower threshold will limit the information loss at the cost of producing more action potentials. Surprisingly, we find that the performance of the spiking network is very close to that of the SPRT, even when  $\tau_s$  is set high, so that a small number of action potentials is produced (see Materials and method for the encoding, Figure 4e through h, Figure 9, and Supplementary Figure S1b through d for the quality of approximation). Since the network behavior is quite insensitive to  $\tau_s$  (see

Supplementary Figure S1), we do not consider  $\tau_s$  to be a free parameter and set its value to  $\tau_s = 0.5$  in our experiments.

### Softmax

One of the fundamental computations in Equation 10 is the softmax function (Equation 6). It requires taking exponentials and logarithms, which have not yet been shown to be within a neuron's repertoire. Fortunately, it has been proposed that softmax may be approximated by a simple maximum (Chen et al., 2011; Ma et al., 2011) and implemented using a winner-take-all mechanism (Koch & Ullman, 1987; Seung, 2009) with spiking neurons (Oster, Douglas, & Liu, 2009). Through numerical experiments we find that this approximation results in almost no change to the network's behavior (see Figure 4e and Supplementary Figure S1a). This suggests that an exact implementation of softmax is not critical, and other mechanisms that may be more neurally plausible have similar performances. The time it takes for the winner-take-all network to converge is typically small (it is on the

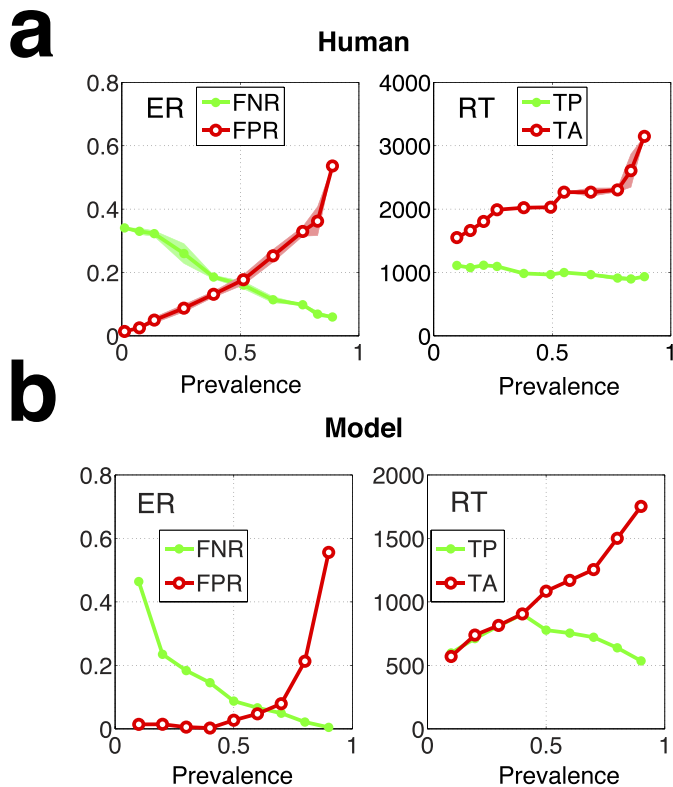


Figure 7. Qualitative predictions of SPRT on the effect of target prevalence (Simulation 5). (a) The median RT (left) and ER (right) of human search performance with varying target prevalence (adapted with permission from figure 2a and c in Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, 20(2), 121–124) by averaging performance of many subjects. (b) Qualitative predictions of SPRT. Simulation parameters are identical to those used in Figure 5.

millisecond level for 10s of neurons and scales logarithmically with the number of neurons; Koch & Ullman, 1987) compared with the interspike intervals of the input neurons (around 30 ms per neuron and 12 ms for a hypercolumn of  $N = 16$  neurons per visual location; Vinje & Gallant, 2000).

### Decision

Finally, the log likelihood ratio  $S(X_t)$  is compared with a pair of thresholds to reach a decision (Equation 3). The positive and negative parts of  $S(X_t)$ ,  $(S(X_t))^+$  and  $(-S(X_t))^+$ , may be represented separately by two mutually inhibiting neurons (Gabbiani & Koch, 1996), where  $(\cdot)^+$  denotes halfwave rectification:  $(x)^+ \triangleq \max(0, x)$ . We can implement Equation 3 by simply setting the firing thresholds of these neurons to the decision threshold  $\tau_1$  and  $-\tau_0$ , respectively.

Alternatively,  $S(X_t)$  may be computed by a mechanism akin to the ramping neural activity observed in decision-implicated areas such as the frontal eye field

(Woodman et al., 2008; Heitz & Schall, 2012; Purcell et al., 2012).  $(S(X_t))^+$  and  $(-S(X_t))^+$  could be converted to two trains of action potentials using the same encoding scheme described in the Signal transduction section. The resultant spike trains may be the input signal of an accumulator model (e.g., Bogacz et al., 2006). The model has been shown to be implementable as a biophysically realistic recurrent network (Wang, 2002; Lo & Wang, 2006; Wong et al., 2007) and capable of producing and thresholding ramping neural activity to trigger motor responses (Mazurek et al., 2003; Woodman et al., 2008; Heitz & Schall, 2012; Purcell et al., 2012; Cassey, Heathcote, & Brown, 2014). While both neural implementations of  $S(X_t)$  are viable options, in the simulations used in this study we opted for the first.

### Network structure

If we combine the mechanisms discussed above (i.e., local gain control, an approximation of softmax, a spike-based coding of analog log likelihood values, and the decision mechanism), we see that the mathematical computations required by SPRT can be implemented by a deep network of spiking neurons (Figure 4a). It comprises local hypercolumn readout networks (Figure 4b) and a central circuit that aggregates information over the visual field. First, every location in the image is analyzed by a heterogeneous discrimination network. This network computes the local log likelihood ratio  $S(X_t^l)$  (Equation 11) and simultaneously computes the local log likelihood for each CDD (Equation 25). The CDD log likelihoods are aggregated over all locations and sent to a gain-control unit, where the posterior of the CDD  $Q_\phi = \log P(\phi|X_t)$  is estimated (see Equation 25). At each time instant this estimate is fed back to the local networks to suppress other CDD estimates and equivalently compute  $S(X_t^l)$  using the best estimate for the set size and orientation difference (Equation 11).

It is important to note that both the structure and the synaptic weights of the visual search network described above were derived analytically from the hypercolumn parameters (the shape of the orientation-tuning curves), the decision thresholds, and the probabilistic description of the task. The network designed for heterogeneous visual search could dynamically switch to simpler tasks by adjusting its priors [e.g.,  $P(\phi)$ ]. The network has only 3 *df* rather than a large number of network parameters (Ma et al., 2011; Krizhevsky, Sutskever, & Hinton, 2012). We discuss this in more detail later.

In conclusion, a close approximation to the SPRT may be implemented by a network of integrate-and-fire neurons and spiking, winner-take-all circuits. The architecture and weights of the network are specified by the probabilistic setup of the search task up to 3 *df*. Next we investigate whether predictions of our model



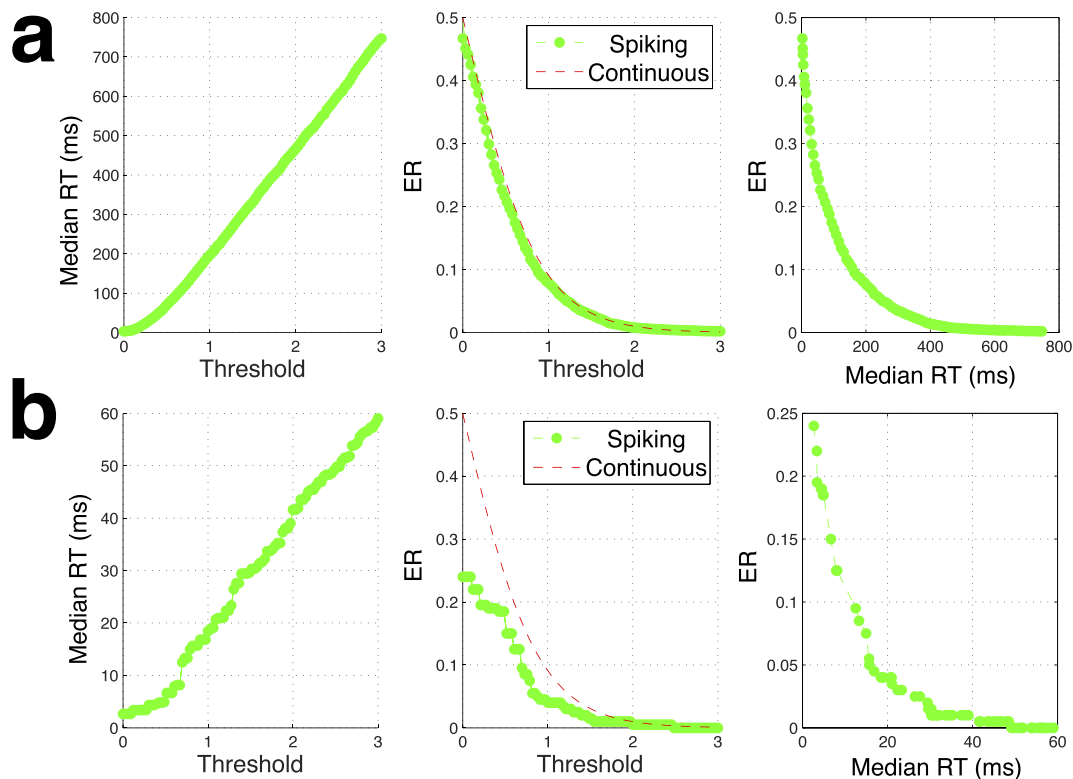


Figure 8. Spiking input causes discontinuous RT–ER tradeoff for easy tasks (Simulation 6). (a) Visual discrimination with an orientation difference of  $15^\circ$  (hard task). Median RT (left) and ER (middle) are plotted with respect to symmetrical decision thresholds and against each other (right). Continuous ER–threshold relationship is computed using formulas in Wald (1945). (b) Visual discrimination with orientation difference of  $90^\circ$  (easy task). Both median RT and ER show discrete jumps as threshold varies smoothly. Actual ERs are lower than those predicted by a continuous model because diffusions vary by discrete action-potential jumps and thus exceed thresholds by a large amount. Simulation parameters are identical to those used in Figure 5.

are consistent with the known literature and assess the optimality of humans in conducting visual search.

## Qualitative predictions

A first test of our model is to explore its qualitative predictions of RT and ER in classical visual search experiments (Figure 1a). In a first simulation experiment (Simulation 1), we used a blocked design (Figure 3a), where the orientation of targets and distractors as well as the number of items do not change from image to image within an experimental block (Figure 5a through c). Thus, the observer knows the value of these parameters from experience. Accordingly, we held these parameters constant in the model. We assume that the costs of error are constant; hence, we hold the decision thresholds constant as well. What changes from trial to trial is the presence and location of the target and the timing of individual action potentials in the simulated hypercolumns.

The model makes three qualitative predictions. First, the RT distribution predicted by the model is heavy tailed: It is approximately log normal in time (Figure

5b). Second, the median RT increases linearly as a function of  $M$ , with a large slope for hard tasks (small orientation difference between target and distractor) and almost no slope for easy tasks (large orientation difference; Figure 5a). Last, the median RT is longer for target absent than for target present, with roughly twice the slope (Figure 5a). The three predictions are in agreement with classical observations in human subjects (Treisman & Gelade, 1980; E. Palmer, Horowitz, Torralba, & Wolfe, 2011).

In a second experiment (Simulation 2) we adopted a mixed design, where the distractors are known but the orientation difference is sampled from  $10^\circ$ ,  $20^\circ$ , and  $60^\circ$ , randomized from image to image. The subjects (and our model) do not know which orientation difference is present before stimulus onset. The predictions of the model are shown in Figure 5c. When the target is present, both RT and ER are sensitive to the orientation difference and will decrease as the orientation difference increases. That is, the model predicts that an observer will trade off errors in difficult trials (more errors) with errors in easy trials (fewer errors) to achieve an overall desired average performance (see Supplementary Figure S5 for the tradeoff per subject),

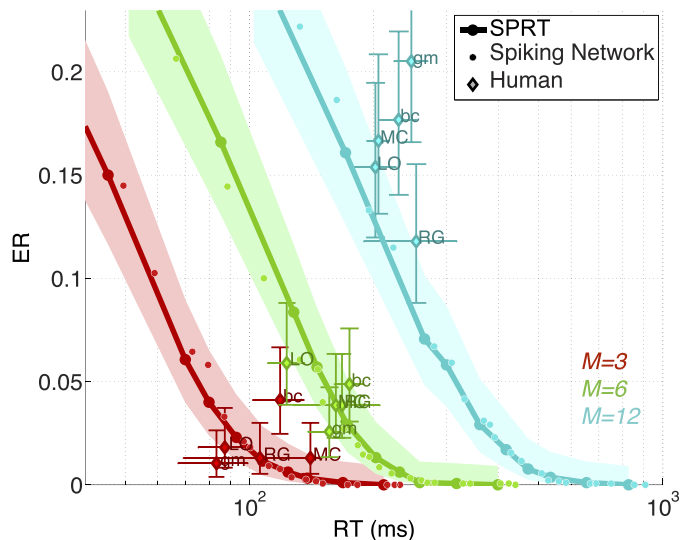


Figure 9. Speed-accuracy tradeoff. ER versus RT tradeoff in the mixed set-size task (Experiment 2; Figure 3c) of five human subjects (G. M., B. C., M. C., L. O., and R. G.) and of SPRT and the spiking network implementation using the same fitted internal parameters of the subjects (see the Psychophysics and Model fitting sections). The set size takes value from {3, 6, 12}, and the orientation difference is fixed at  $30^\circ$ . SPRT uses the generative model shown in Figure 1b.

which is consistent with the psychophysics data (see the Psychophysics section).

In a third experiment (Simulation 3) we used items with four orientations ( $\theta = 0^\circ, 12^\circ, 25^\circ$ , and  $37^\circ$ ; see Figure 5c) and explored two conditions: one in which the orientation of the targets is larger than that of the distractors ( $0^\circ, 12^\circ, 25^\circ$ , and  $37^\circ$ ; targets in bold) and one in which the orientations are interleaved ( $0^\circ, 12^\circ, 25^\circ$ , and  $37^\circ$ ). Our model predicts that the difficulty of the task (median RT at given ER) is much higher in the second condition, even if the minimum orientation difference between the target and the distractor set is the same. This observation matches observations in the psychophysics literature (Duncan & Humphreys, 1989; Hodsoll & Humphreys, 2001).

In Simulation 4 we explored which of two competing models best accounts for visual search when scene complexity is unknown in advance. Recall that in the Heterogeneous search section we proposed two models: the SPRT (Figure 4a), which is near optimal, and a simplified model (Equation 28), which is suboptimal. SPRT estimates the scene complexity parameter trial by trial (Equation 10) and predicts that ERs are comparable for different set sizes, whereas RTs show strong dependency on set size when the orientation difference is small (Figure 6a). The simplified model, in which scene complexity is assumed to be constant (Equation 28), predicts the opposite—that ER depends strongly on set size, whereas RT is almost constant when the

target is present (Figure 6b). Human psychophysics data (Wolfe et al., 2010; reproduced in Figure 6c) show a positive correlation between RT and set size and little dependency of ER on set size, which favors SPRT and suggests that the human visual system estimates scene complexity while it carries out visual search.

In Simulation 5, target frequency (commonly called *target prevalence*) is varied systematically (Wolfe & Van Wert, 2010). The model's prediction (Figure 7b) matches qualitatively human psychophysics (Wolfe & Van Wert, 2010; reproduced in Figure 7a). Changing target frequency results in a more pronounced change in target-absent RTs than in target-present RTs. False-negative rate is negatively correlated with target prevalence, whereas false-positive rate is the opposite.

In a last simulation experiment (Simulation 6), we explored the effect of spike quantization on ER and RT. We conjectured that one ought to see quantization effects in both ER and RT for easy tasks. Consider the visual discrimination task where only one object, either a target or a distractor, is displayed. This is equivalent to a search task with  $M = L = 1$ . When the orientation difference is  $90^\circ$ , most neurons in the hypercolumn can easily discriminate the target from the distractor. As a result, most action potentials will cause big jumps in the diffusion (Equation 12), and a decision may be made after observing very few action potentials. For example, after one or two spikes, the log likelihood will most likely be either above 0.5 or below  $-0.5$ . Therefore, any threshold in  $(0, 0.5]$  would achieve the same effect as the threshold  $\tau = 0.5$ , which corresponds to an ER of 24% (we assume  $-\tau^0 = \tau^1 = \tau$ ). Indeed, our model predicts that the ER will be either 50% or less than 24% (Figure 8b). Furthermore, the model predicts that ERs around 20% to 25% would be more frequently observed than ERs around 10% to 15%. This quantization effect continues and gradually dissipates as the threshold is increased because more spikes are needed for the log likelihood to cross the threshold. We do not find in the literature any study describing this phenomenon. We can only assume that such an observation would not be considered worth reporting in the absence of a suitable theoretical framework.

## Psychophysics

In order to assess our model quantitatively, we compared its predictions with data harvested from human observers who were engaged in visual search (Figure 1a). Three experiments were conducted to test both the model and humans under different conditions. The conditions are parameterized by the orientation difference chosen from  $\{20^\circ, 30^\circ, 45^\circ\}$  and the set size from  $\{3, 6, 12\}$ . The blocked design was used in the first experiment (Experiment 1), where all nine ( $3 \times 3$ ) pairs

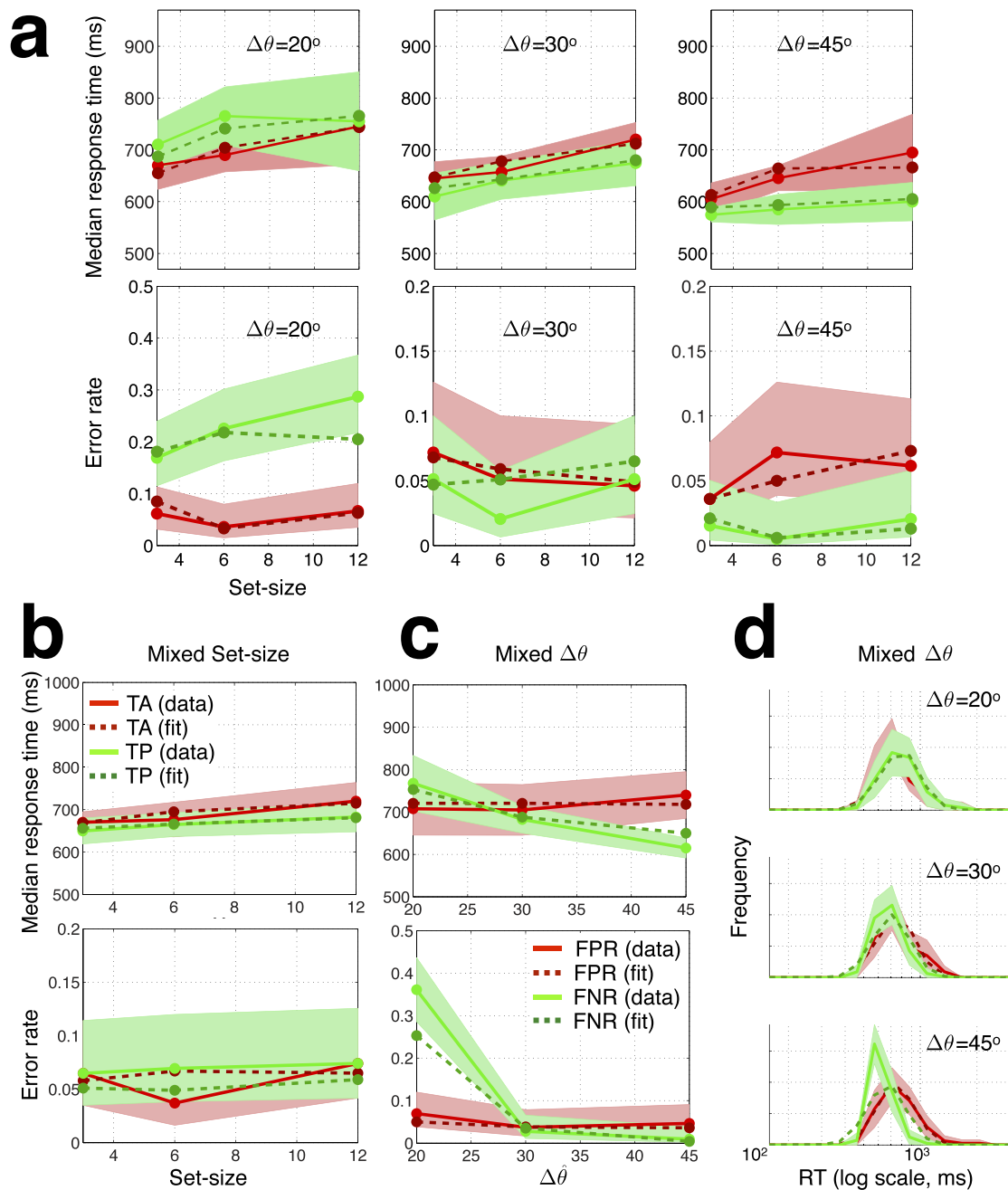


Figure 10. Behavioral data of a randomly selected human subject and fits (ER, median RT, and RT distributions) using SPRT. (a) Experiment 1: the blocked design. All set size  $M$  and orientation difference combinations share the same hypercolumn and nonperceptual parameters; the decision thresholds are specific to each  $\Delta\theta$ – $M$  pair. Fits are shown for RTs (first row) and ER (second row). (b–c) RT and ER for (b) Experiment 2, the mixed set size, and (c) Experiment 3, the mixed contrast ( $\Delta\theta$ ) design. (d) RT histogram for the mixed  $\Delta\theta$  design, grouped by orientation difference. For each design all combinations share the same thresholds, hypercolumn, and nonperceptual parameters. Color indicates target present (TP) or target absent (TA); solid lines represent data, and dashed lines represent model (see panel c). See the data and fits of all individual subjects in Supplementary Figures S3, S4, and S5.

of orientation difference and set size combinations were tested in blocks. The second experiment randomized the set size while holding the orientation difference fixed at  $30^\circ$  (Experiment 2). The third randomized orientation difference from trial to trial while fixing the set size at 12 (Experiment 3). The subjects were

instructed to maintain eye fixation at all times and respond as quickly as possible and were rewarded based on accuracy (see the Psychophysics procedure section).

We fit our model to explain the full RT distributions and ERs for each design separately. In order to

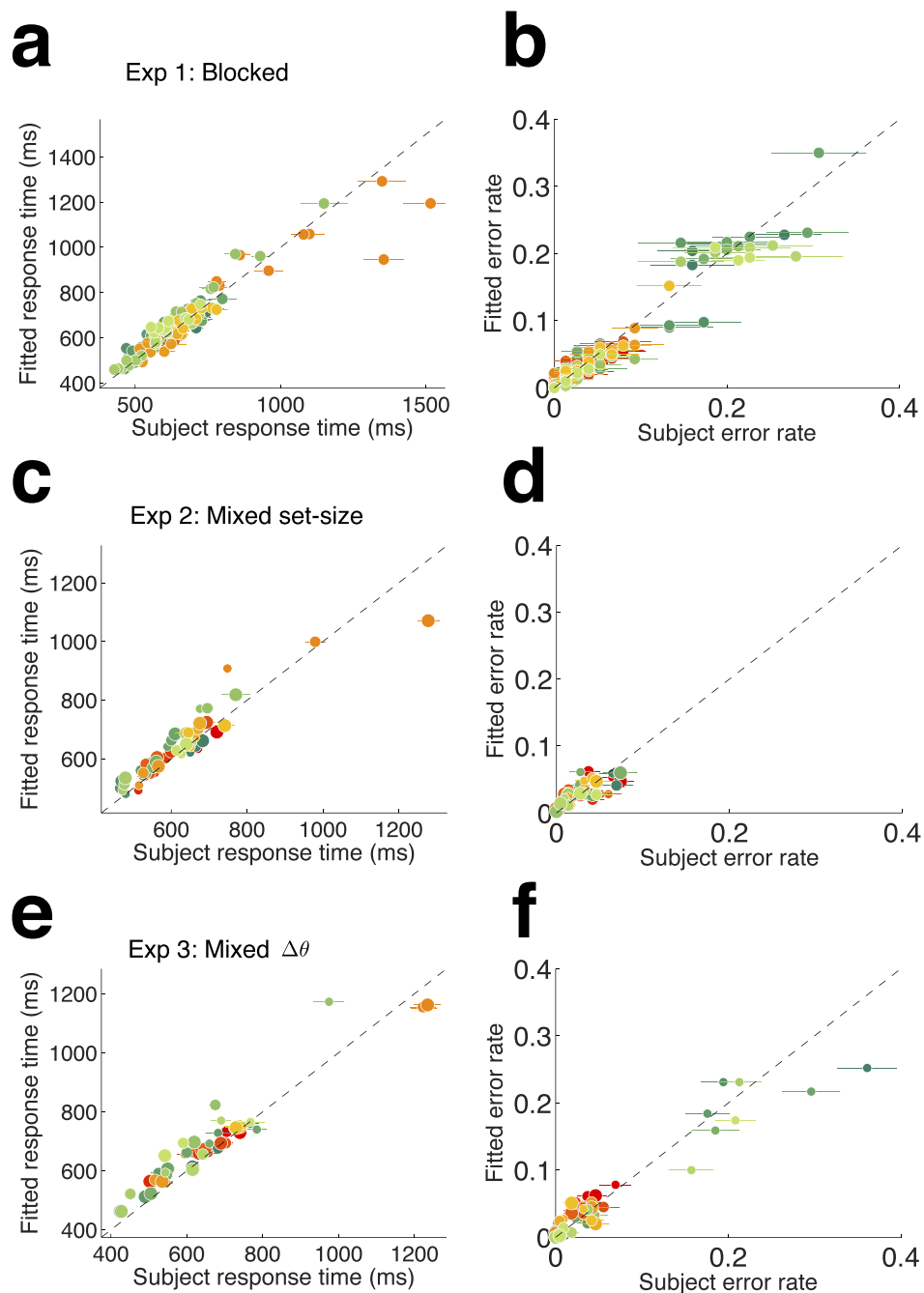


Figure 11. Synopsis of fits to nine individual subjects. The rows correspond respectively to three designs: Experiment 1 (blocked), Experiment 2 (mixed set size), and Experiment 3 (mixed contrast;  $\Delta\theta$ ). The columns correspond to median RT and ER. Each plot displays fitted values against the data (the perfect fit is indicated by the dashed line). The maximum firing rate of the hypercolumn  $\lambda_{\max}$  and the two nondecision parameters for each subject are fitted using only the blocked design experiment and used to predict median RT and ER for the two mixed design experiments. Error bar shows 1 SE of the data. Colors are specific to subject. The small, medium, and large dots correspond respectively to the orientation differences 20°, 30°, and 45° in panels c and d and to the set sizes 3, 6, and 12 in panels e and f.

minimize the number of free parameters, we held the number of hypercolumn neurons constant at  $N = 16$  (see Discussion for the plausibility of  $N$ ), their minimum firing rate constant at  $\lambda_{\min} = 1$  Hz, and the half-width of their orientation tuning curves at 22° (full

width at half height = 52°; Graf et al., 2011). Hence, we were left with only three free parameters: The maximum firing rate of any orientation-selective neuron  $\lambda_{\max}$  controls the signal-to-noise ratio of the hypercolumn, and the upper and lower decision



thresholds  $\tau_0$  and  $\tau_1$  control the frequency of false-alarm and false-reject errors. Once these parameters are given, all the other parameters of our model are analytically derived.

Two additional free parameters were required to fit our subjects' data. It is well known that the RT consists of a perceptual component and a nonperceptual motor and neural conduction component (E. Palmer et al., 2011). The perceptual component is the time for a search decision to be reached and is accounted for by our model with three free parameters, as explained in the previous paragraph. The nonperceptual delay accounts for axonal conduction, muscle activation, and other factors external to visual search. We modeled the latter phenomenologically with a log-normal distribution parameterized by two parameters: mean  $\mu_D$  and log-time variance  $\sigma_D^2$ . Therefore, we fit RT distributions and ERs with five parameters (three for SPRT and two for nonperceptual delay).

In blocked-design Experiment 1, the hypercolumn and the motor time parameters were fit jointly across all blocks (about 1,620 trials); the decision thresholds were fit independently on each block (180 trials/block). Since the degree of difficulty was different for each block and a subject's costs of errors  $C_p$  and  $C_n$  might also vary for each block, we could not assume that the subject's thresholds would remain constant. Therefore, in blocked design experiments, 21 parameters (2 thresholds  $\times$  9 conditions + 1 SNR + 2 motor parameters) were used to fit nine conditions, each containing 180 target-present trials and 180 target-absent trials. In mixed-design Experiments 2 and 3, all five parameters were fit jointly across all conditions for each subject because all conditions are mixed. Thus, five parameters (which was reduced to two in the generalization experiment below) were used to fit three conditions, each containing 220 target-present trials and 220 target-absent trials (see Equation 16 for the fitting procedure; Figure 10 for data and fits of a randomly selected individual; Supplementary Figures S3, S4, and S5 for data and fits for every subject; Figure 9 for the ER vs. RT tradeoff curve fit to five subjects with similar signal-to-noise ratio; and Figure 11a and b for all subjects in the blocked experiment).

In each experiment the model is able to fit the subjects' data well. The parameters that the model estimated (the maximum firing rate of the neurons  $\lambda_{\max}$  and the decision thresholds  $\tau_0$ ,  $\tau_1$ ) are plausible (Vinje & Gallant, 2000; see Discussion for the plausibility of  $\lambda_{\max}$ ). Subjects had similar parameters, although intersubject variability is noticeable (see Supplementary Figures S3, S4, and S5). Each subject displays different ERs for different conditions (see Figure 9); thus, the decision thresholds are indeed not constant (see Supplementary Figures S3, S4, and S5 for fitted thresholds). It may be possible to model the inter-

condition variability of the thresholds as the result of the subjects minimizing a global risk function (Drugowitsch et al., 2012). Therefore, for each subject in blocked-design Experiment 1, we have tried fitting a common Bayes risk function (Equation 2), parameterized by the two costs of errors  $C_p$  and  $C_n$ , across all blocks and solving for the optimal thresholds for each block independently. This assumption reduces the number of free parameters from 21 to five (2 costs of errors + 1 SNR + 2 motor parameters), and it leads to marked reduction in the quality of fits for some of the subjects (see Supplementary Figure S6). Therefore, as far as our model is concerned, there was some block-to-block variability of the error costs.

Finally, we test the generalization ability of our model. We used the signal-to-noise ratio parameter (the maximum firing rate  $\lambda_{\max}$ ) and the two nondecision delay parameters estimated from the blocked experiment (Experiment 1) to predict the mixed experiments (Experiments 2 and 3). Thus, for each mixed experiment only two parameters—namely the decision thresholds  $\tau_0$  and  $\tau_1$ —were fit. Despite the parsimony in parameterization, the model shows good cross-experiment fits (see Figure 11c through f), suggesting that the parameters of the model refer to real characteristics of the subject.

## Discussion

Searching for objects among clutter is one of the most valuable functions of the sensory systems. Best performance is achieved with fast RT and low ERs; however, RT and ERs are competing requirements that have to be traded off against each other. The faster one wishes to respond, the more errors one makes due to the limited rate at which information flows through the senses. Conversely, if one wishes to reduce ERs, decision times become longer. In order to study the nature of this tradeoff we derived a near-optimal decision strategy for visual search. The input signal to the model is action potentials from orientation-selective hypercolumn neurons in the primate striate cortex V1, and the output of the model is a binary decision (target present vs. target absent) and a decision time.

Five free parameters uniquely characterize the model: the maximum firing rate of the input neurons, the maximum tolerable false-alarm and false-reject ERs, and two parameters characterizing response delays that are unrelated to decision. Once these parameters are set, RT histograms and ER may be computed for any experimental condition. Our model may be implemented by a deep neural network comprising five layers of neuron-like mechanisms. Signals propagate from layer to layer mostly in a feed-

forward fashion; however, we find that two feedback mechanisms are necessary: (a) gain control (lateral inhibition) that is local to each hypercolumn and has the function of maintaining signals within a small dynamic range and (b) global inhibition that estimates the complexity of the scene. To the best of our knowledge a mechanism that estimates the complexity of the scene had not been postulated and tested before. Comparison of model predictions with human behavior suggests that the visual system of human observers indeed does estimate scene complexity as it carries out visual search and that this estimate is used to control the gain of decision mechanisms.

The structure of our model is completely determined by the system of equations (Equations 10 and 11). This system of equations naturally suggests the five-layer architecture that is shown in Figure 4a and b. However, the architecture is not unique in carrying out the necessary computations. It is well known that any function that is implementable by a multilayer network may be implemented by an appropriately designed two-layer network (Cybenko, 1989). We prefer a five-layer network because it is exponentially more space economical (explained below; Bengio, 2009). For example, to model heterogeneous visual search, a three-layer network requires an exponential number of neurons in the display size  $L$ , whereas a five-layer or deeper network requires only a linear number. For  $L = 24$  locations and  $n_T = n_D = 3$ , the three-layer network requires at least  $7 \times 10^{12}$  neurons, whereas the five-layer network needs less than 1,000. We do not add a sixth layer to marginalize out the CDD but instead use a gain-control circuit in parallel to the five-layer network (see Figure 4a and Equation 25). This design is such that the parallel circuit may be easily shared in other tasks where scene complexity and/or orientation difference estimation is necessary. As a result, although alternative architectures are consistent with both our analysis and the data, the five-layer architecture appears to be the better choice.

We explore two versions of the network. In the first version, which implements SPRT exactly, the signals that propagate from layer to layer are analog and represent log likelihood ratios. In the second version, all information propagates by means of action potentials. A surprising finding is that even when the firing rate is limited to about 10 action potentials per second, the spiking model produces an RT-versus-ER curve that is very close to that produced by the optimal analog model. Since the spiking model may be implemented by plausible neuronal mechanisms, the computations needed to implement SPRT could be carried out in the cortex.

We collected RTs and ERs from 10 subjects who carried out a set of diverse visual search tasks; we fit our model to each subject individually. We find that the

model fits a variety of conditions well despite having only three free parameters in addition to two motor-response parameters. Our model uses  $N = 16$  uncorrelated, orientation-tuning neurons per visual location, each with a half tuning width of  $22^\circ$  and a maximum firing rate (estimated from the subjects) of approximately 17 Hz. The tuning width agrees with V1 physiology in primates (Graf et al., 2011). Although our model appears to have underestimated the maximum firing rate of cortical neurons (which ranges from 30 to 70 Hz; Graf et al., 2011) and the population size  $N$  (which may be in the order of hundreds), actual V1 neurons are correlated; hence, the equivalent number of independent neurons is smaller than the measured number. For example, take a population of  $N = 16$  independent Poisson neurons, all with a maximum firing rate of 17 Hz, and combine every three of them into a new neuron. This will generate a population of 560 correlated neurons with a maximum firing rate of 51 Hz and a correlation coefficient of 0.19, which is close to the experimentally measured average of 0.17 (Graf et al., 2011; see Vinje & Gallant, 2000, for a detailed discussion on the effect of sparseness and correlation between neurons). Therefore, our estimates of the model parameters are consistent with primate cortical parameters. The parameters of different subjects are close but not identical, matching the known variability within the human population (Van Essen, Newsome, & Maunsell, 1984; E. Palmer et al., 2011). Finally, the fact that estimating model parameters from data collected in the blocked experiments allows the model to predict data collected in the mixed experiments does suggest that the model parameters mirror physiological parameters in our subjects.

What is the relationship between our model and an ideal observer? An ideal observer is a system that characterizes the optimal performance given the available information and specific constraints (Geisler, 2011). In our model the signal goes through two stages: (a) an input stage, which transforms images into cortical activity patterns, and (b) a decision stage, which reads cortical activity and computes a decision. The second stage is near optimal; that is, given the signal produced by the first stage, the second stage produces decisions that are indistinguishable from optimal.

The first stage is instead not optimal. To start with, the visual system discards photons at the outset by various gain-control mechanisms in the eye. Furthermore, the firing rate of cortical neurons is limited by physiology, and their total number is limited by anatomy. The performance of the system is thus limited by anatomical and physiological constraints. An ideal observer that does not take these constraints into account will seriously overestimate the performance of the visual system. Therefore, in modeling visual search

it is inevitable to treat the early visual system (eye to cortex) as a computational bottleneck with parameters that need to be adjusted for each subject.

We chose to model the cortex as an idealized collection of neurons whose firing rate is Poisson and, conditioned on the stimulus, independent of each other (the LNP model; Chichilnisky, 2001; Simoncelli et al., 2004). This model has a number of advantages: It is simple and parsimonious, it is well studied in the literature (Goris, Movshon, & Simoncelli, 2014), and its limitations are increasingly well understood (Goris et al., 2014).

Is our model an ideal observer? From the discussion above, it is clear that the short answer is *no*. The long answer is that our model is an ideal observer if one agrees to use the LNP model to capture the computational limitations that are imposed by specific assumptions on the biophysics, physiology, and anatomy of the cortex—most specifically, the firing rate of cortical neurons, the number of neurons in a hypercolumn, and the tuning width of orientation-specific neurons. We can estimate these parameters by fitting the experimental data, but to address the optimality question one would need to either measure these quantities directly in human subjects or carry out the psychophysics in laboratory primates in which these numbers are known. The second route appears to be more practical.

Despite this limitation, valuable insight may be gained from our model. First, the fact that the model qualitatively fits and predicts complex data that were collected by testing subjects in a variety of different conditions suggests that the model may capture the nature of the computations that are carried out by the visual system during visual search. The parsimony in the number of parameters makes this particularly compelling. Second, the model produces first-existence proof that such computations may be carried out by maps of cortical spiking neurons (i.e., by known mechanisms that are well documented in the physiology literature). Third, the estimates of the parameters of the model (i.e., number of neurons, tuning width, firing rates) that we obtain by fitting our observer's data are in line with the estimates that come from model primates, lending further plausibility to the model.

Our model was developed under several simplifying assumptions. First, we used simple stimulus patterns comprising oriented bars following traditions in psychophysics. More complex stimuli would require augmenting the front end of the model with additional neuronal populations. For example, our model could be adapted to search for color if appropriate color-selective neurons were added to the input layer of the model. Second, we sidestepped both retinal and cortical magnification (DeValois & DeValois, 1980) by placing items on a constant-eccentricity ring (Figure 1a). Accounting for magnification as a function of eccentricity is not conceptually difficult, but it would

introduce more free parameters and unnecessarily complicate the model. Third, our model does not account for attention (Treisman & Gelade, 1980; Desimone & Duncan, 1995; Wolfe & Horowitz, 2004). We eliminated overt attention in the psychophysics experiments by enforcing eye fixation at all times. Covert shifts of attention (Sperling & Melchner, 1978; Posner, Cohen, & Rafal, 1982; Moore & Fallah, 2004) are unnecessary to explain the variability of ER and RT with respect to scene complexity as may be seen in our qualitative predictions and as explained by signal detection theory (Green & Swets, 1966; J. Palmer et al., 2000). Extending our model to include optimal eye-movement planning (Najemnik & Geisler, 2005; Rutishauser & Koch, 2007) is an interesting direction that goes beyond the scope of this work.

Our near-optimal model does not use a diffuse-to-bound mechanism. Instead, local diffusions are combined using nonlinear functions (softmaxes). Given task parameters (set sizes, orientation differences, and the uncertainty associated with each), the construction of the model is parameter free. By contrast, a diffusion model may require additional parameters specifying how the statistics of the diffusions relate to the task parameters (J. Palmer et al., 2005; Drugowitsch et al., 2012). Moreover, Figure 11c through f showcases our model's ability to generalize to novel experimental settings, which is nontrivial for diffusion models. Furthermore, our model connects the underlying cortical mechanisms and physiological parameters to the subjects' behavior. For example, the prediction that the visual system employs gain-control mechanisms for estimating the complexity of the scene would not be possible if one just fit the data with diffusions. Therefore, while diffusion models remain great phenomenological models for decision-making mechanisms, we use the optimal decision strategy to study the optimal tradeoff of ER and RT across different experimental settings.

Finally, our spiking model suggests that behavior arising from action potentials may be quantized (Perona, 2014). In an easy task where the observer is pressured to respond extremely quickly, only a few action potentials are necessary to trigger a decision. In that case our model predicts a quantized ER as the decision threshold varies smoothly. Observing the quantized ER experimentally would allow us to reveal decisions that are based on very few action potentials.

## Materials and method

### Ethics statement

This study was performed in strict accordance with the ethical principles set forth in the Declaration of



Helsinki and the report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The protocol was approved by the Caltech Institutional Review Board (Protocol Title: Measuring Response Time and Error Rates in Visual Search, Number: 13-0374). Informed consent was obtained from all participants.

## Psychophysics procedure

We asked 10 subjects (aged 18–23 years) with normal or corrected vision to search visually for a target in images that were presented on a screen. Each trial began with a blank screen with a cross in the middle, which the subjects were instructed to fixate. This fixation screen was displayed for a uniformly distributed random duration of 500 to 1200 ms. After the fixation screen, a stimulus image appeared and remained on the screen until subjects responded (Figure 1a). Subjects were instructed to indicate accurately and as quickly as possible whether a target was present or absent by pressing one of two keyboard keys. Visual feedback was given for 500 ms immediately after the response, concluding the trial. In order to motivate subjects to be more accurate, subjects were paid based on their average accuracy. Pay for a day's work (approximately 1 hr) ranged from a minimum of \$5 when ER was greater than 35% to a maximum of \$20 when ER was less than 5%. For intermediate ERs pay was pro rata (i.e., 50¢ for every 1% improvement in ER). One subject's data were excluded from analysis due to extremely poor accuracy.

Images were displayed on a 24-in. light-emitting diode monitor with a 60-Hz refresh rate viewed at a distance of 50 cm. Each image comprised  $0.11^\circ \times 0.26^\circ$  (width  $\times$  length) oriented dark bars, evenly spaced along the arc of a circle  $16^\circ$  in radius and presented against a uniform bright background. The spacing between consecutive bars was one sixteenth of the perimeter of the full circle (Figure 1a). In each presentation, the arc's starting angle was randomized and the  $x$ - $y$  location of each bar was jittered to prevent crowding. The random jittering vector for each bar alternated between pointing inward and outward; its magnitude was chosen uniformly at random from  $[0, 0.6^\circ]$  and its orientation was chosen from  $[-5.6^\circ, 5.6^\circ]$ . Unless otherwise specified, distractors were oriented at  $30^\circ$  orientation difference (the difference in orientation between target and distractor bars) and set size (the total number of bars in the image) was systematically varied in our experiments. Orientation difference was chosen from  $\{20^\circ, 30^\circ, 45^\circ\}$  and set size was chosen from  $\{3, 6, 12\}$ . Targets were present in 50% of the images in random order.

Our subjects' performance was measured with three different experimental designs. In Experiment 1, subjects saw a block of 60 trials (each trial corresponds to one image stimulus), with orientation difference and set size held constant throughout each block and varied across blocks. Blocks were organized in sessions, where each session contains nine different blocks covering all orientation difference and set size combinations. In Experiment 2, subjects saw a session of 240 trials, with set size varied randomly from trial to trial and orientation difference fixed at  $30^\circ$ . Eighty trials of each set size were presented in each session. In Experiment 3, the set size was held constant at 12 items and orientation difference was varied randomly within the session. Again, 80 trials were presented for each orientation difference.

Subjects were trained for 7 days and tested for 3 days (approximately 1 hr/day). Each training day consisted of two sessions of each of the three experiments. Each test day consisted of eight sessions of a single experiment. Only data from the test days are presented here. We analyzed the data collected during the training days to verify that the subjects' performance had stabilized.

To ensure eye fixation, we randomly selected one subject to perform 1 day of the experiment while being monitored by an eye tracker. The subject showed no significant eye movement or difference in performance.

## Model fitting

### Hypercolumns

Our V1/V2 front end consists of hypercolumns, each comprising  $N$  orientation-tuned neurons whose preferred orientations are distributed uniformly over the  $[0^\circ, 180^\circ]$  interval. The tuning curves, given in Equation 1, are parameterized by the minimum and maximum firing rates  $\lambda_{\min}$  and  $\lambda_{\max}$  as well as the half tuning width  $\psi$ .  $\lambda_{\min}$  was fixed at one spike per second (Graf et al., 2011). Since increasing  $\lambda_{\max}$  and the number  $N$  of neurons achieves the same effect of boosting the neuron's signal-to-noise ratio, we fixed  $N = 16$  and varied  $\lambda_{\max}$  only. The half tuning width of the neurons was set to  $\psi = 22^\circ$  for all subjects. Therefore, the only tuning parameter for the front end is  $\lambda_{\max}$ .

### Decisions

The model computes present and absent decisions by comparing the log probability ratio  $S(X_t) = \log \frac{P(C=1|X_t)}{P(C=0|X_t)}$  (Equation 10) with thresholds  $\tau_0$  and  $\tau_1$ . As soon as either one of the two threshold is exceeded, the corresponding decision is taken. This is equivalent to thresholding the probabilities  $P(C=1|X_t)$  and  $P(C=0|X_t)$  with thresholds  $P_1$  and  $P_0$  since the probabilities



and the likelihood ratio are related by the expression

$$\begin{aligned} P(C = 1|X_t) &= \frac{P(C = 1|X_t)}{P(C = 1|X_t) + P(C = 0|X_t)} \\ &= \frac{1}{1 + \exp\left(-\log \frac{P(C=1|X_t)}{P(C=0|X_t)}\right)} = h(S(X_t)) \end{aligned} \quad (13)$$

$$P_0 = h(\tau_0), \quad P_1 = h(\tau_1), \quad (14)$$

where  $h(x)$  is the logistic function, which is monotonically increasing, and where  $P(C = 0|X_t) = 1 - P(C = 1|X_t)$ . Therefore, testing whether  $S(X_t) > \tau_1$  is equivalent to testing whether  $P(C = 1|X_t) > P_1 = h(\tau_1)$  and similarly testing whether  $S(X_t) < \tau_0$  is equivalent to testing  $P(C = 0|X_t) < P_0$ . The thresholds  $\tau_0$  and  $\tau_1$  are additional free parameters in the model.

### Nonperceptual response delay

Our subjects' RTs are the sum of the perceptual RT  $T_p$ , which our model predicts, and additional delays  $T_m$  due to axonal transmission, muscle activation, and other factors external to the visual search process (Schwarz, 2001; E. Palmer et al., 2011). We model the statistics of the nonperceptual response delay with a log-normal distribution, a more realistic model than the exponential (Schwarz, 2001). The log-normal is parameterized by its mean  $\mu_D$  and variance  $\sigma_D$ , which are fitted separately for each subject.

### Parametrization of the model

In summary, the model has five free parameters. Parameters  $\lambda_{\max}$ ,  $\mu_D$ , and  $\sigma_D$  are specific to each subject, and the decision thresholds ( $\tau_0$  and  $\tau_1$ ) are specific to each subject in each task condition. We fit parameters ( $\lambda_{\max}$ ,  $\mu_D$ ,  $\sigma_D$ ) for each subject and ( $\tau_0$  and  $\tau_1$ ) for each block in blocked conditions and for each of the mixed conditions.

### Parameter fitting

The parameters were estimated using maximum likelihood (details below). Figure 11 shows fits for all subjects in all conditions; individual fits are shown in Figure 10 and Supplementary Figures S3, S4, and S5.

The maximum likelihood procedure searches for the model parameters that maximize the likelihood of the data. Observations are collected from each subject for each orientation difference  $\Delta\theta$ , set size  $M$ , and stimulus class  $C$ —collectively denoted as the *experimental condition* ( $\theta_{\text{cond}}$ ). Each observation is a pair consisting of the RT  $t_i$  and decision  $d_i \in \{0, 1\}$ . The data set for each experimental condition is  $\mathcal{D}(\theta_{\text{cond}}) \triangleq \{t_i, d_i\}$ . Given

a set of parameters  $\theta_{\text{model}} = \{\lambda_{\max}, \tau_0, \tau_1, \mu_D, \sigma_D\}$  and a condition  $\theta_{\text{cond}}$ , the Bayesian model computes the perceptual RT distribution  $T_p \sim P_{T_p}(\cdot|\theta_{\text{cond}}, \theta_{\text{model}})$  and the rate  $\alpha(\theta_{\text{cond}}, \theta_{\text{model}})$ . The occurrence of an error trial thus follows a Bernoulli probability with mean  $\alpha(\theta_{\text{cond}}, \theta_{\text{model}})$ . Recall that the total RT  $T$  is modeled as the sum of two variables—the perception time variable  $T_p$  simulated from the Bayesian observer and the log-normally distributed, nonperceptual motor and propagation delay  $T_m$ :

$$\begin{aligned} P_T(T = \tau|\theta_{\text{cond}}, \theta_{\text{model}}) &= \int_t P_{T_p}(T_p = t|\theta_{\text{cond}}, \lambda_{\max}, \tau_0, \tau_1) \\ &\quad \times \log\mathcal{N}(T_m = \tau - t|\mu_D, \sigma_D) dt. \end{aligned} \quad (15)$$

Finally, the subjects' response in each trial is assumed to be independent from the response in other trials. Thus, the likelihood of a set of observations  $\mathcal{D}(\theta_{\text{cond}})$  is given by

$$\begin{aligned} P(\mathcal{D}(\theta_{\text{cond}})|\theta_{\text{model}}) &= \prod_i P_T(t_i|\theta_{\text{cond}}, \theta_{\text{model}}) \\ &\quad \times B(\mathbb{I}(d_i \neq C)|\alpha(\theta_{\text{cond}}, \theta_{\text{model}})), \end{aligned} \quad (16)$$

where  $B(\cdot|\beta)$  is the Bernoulli distribution with mean  $\beta$ , and  $\mathbb{I}(\text{event})$  is 1 when the event is true and 0 otherwise. In order to estimate the parameters  $\theta_{\text{model}}$  given a set of observations, we sample the space of parameters, compute the likelihood of each set of parameters using Equation 16, and select the parameters with the highest likelihood.

### Orientation log likelihood $\mathcal{L}_\theta$

At each location SPRT computes the log likelihood for each task-relevant orientation from evidence  $X_t$  (in this section we are concerned with one location only; therefore, we omit the location superscript  $l$  to simplify notation), which is a set of spike trains from  $N$  orientation-tuned neurons (which can be generalized to be sensitive to color, intensity, and so on) collected during the time interval  $(0, t)$ . Let  $X_t^{(i)}$  be the set of spikes from neuron  $i$  in the time interval from 0 to  $t$ ,  $K_t^i$  the number of spikes from neuron  $i$  in  $X_t^i$ , and  $K_t$  the total number of spikes. Then the likelihood of  $X_t^{(i)}$  when stimulus orientation is  $\theta$  is given by a Poisson distribution:

$$P(X_t^{(i)}|Y = \theta) = \text{Poiss}(K_t^i|\lambda_\theta^i t) = (\lambda_\theta^i t)^{K_t^i} \frac{\exp(-\lambda_\theta^i t)}{K_t^i!}, \quad (17)$$

where  $\lambda_\theta^i$  is the firing rate of neuron  $i$  when the stimulus orientation is  $\theta$ .

The observations from the hypercolumn neurons are independent from each other. Thus, the log likelihood of  $X_t$  is given by

$$\begin{aligned}\mathcal{L}_\theta(X_t) &\triangleq \log P(X_t | Y = \theta) \\ &= \log \prod_{i=1}^N P(X_t^{(i)} | Y = \theta) \\ &= \sum_{i=1}^N \log \left( (\lambda_\theta^i t)^{K_i} \frac{\exp(-\lambda_\theta^i t)}{K_i!} \right) \\ &= \sum_{s=1}^{K_t} W_\theta^{i(s)} - t \sum_{i=1}^N \lambda_\theta^i + \text{const},\end{aligned}\quad (18)$$

where  $W_\theta^i = \log \lambda_\theta^i$  is the contribution of each action potential from neuron  $i$  to the log likelihood of orientation  $\theta$ , and const is a term that does not depend on  $\theta$  and is therefore irrelevant for the decision. The first term is the diffusion that introduces jumps in  $\mathcal{L}_\theta(X_t)$  whenever a spike occurs. The second term is a drift term that moves  $\mathcal{L}_\theta(X_t)$  gradually in time. When the tuning curves of the neurons regularly tessellate the circle of orientations, as is the case in our model (Figure 4c), the average firing rate of the hypercolumn under different orientations is approximately the same (Figure 4d) and the drift term may be safely omitted from models.

## Review: Bayesian inference for discrimination and homogeneous search

We first rederive the log likelihood ratio  $S(X_t)$  for visual discrimination. For all derivations below we show how to compute  $\log \frac{P(X_t|C=1)}{P(X_t|C=0)}$  from the orientation log likelihoods  $\mathcal{L}_\theta(X_t)$ , keeping in mind that

$$\begin{aligned}S(X_t) &= \log \frac{P(C=1|X_t)}{P(C=0|X_t)} \\ &= \log \frac{P(X_t|C=1)}{P(X_t|C=0)} + \log \frac{P(C=1)}{P(C=0)}.\end{aligned}$$

In homogeneous discrimination, the target and distractor have distinct and unique orientations  $\theta_T$  and  $\theta_D$ . Therefore,

$$\begin{aligned}\log \frac{P(X_t|C=1)}{P(X_t|C=0)} &= \log \frac{P(X_t|\theta=\theta_T)}{P(X_t|\theta=\theta_D)} \\ &= \mathcal{L}_{\theta_T}(X_t) - \mathcal{L}_{\theta_D}(X_t),\end{aligned}\quad (19)$$

which proves Equation 4.

In heterogeneous discrimination,  $\theta_T \in \Theta_T$  and  $\theta_D \in \Theta_D$ . For simplicity assume uniform prior on both target and distractor orientation; that is,  $P(\theta|C=1) = 1/n_T$ ,  $\forall \theta \in \Theta_T$ , and  $P(\theta|C=0) = 1/n_D$ ,  $\forall \theta \in \Theta_D$ :

$$\begin{aligned}&\log \frac{P(X_t|C=1)}{P(X_t|C=0)} \\ &= \log \frac{P(X_t|\theta \in \Theta_T)}{P(X_t|\theta \in \Theta_D)} \\ &= \log \left( \sum_{\theta \in \Theta_T} P(X_t|\theta) P(\theta|C=1) \right) \\ &\quad - \log \left( \sum_{\theta \in \Theta_D} P(X_t|\theta) P(\theta|C=0) \right) \\ &= \log \left( \sum_{\theta \in \Theta_T} \frac{\exp(\mathcal{L}_\theta(X_t))}{n_T} \right) \\ &\quad - \log \left( \sum_{\theta \in \Theta_D} \frac{\exp(\mathcal{L}_\theta(X_t))}{n_D} \right) \\ &= S_{\max_{\theta \in \Theta_T}} \left( \mathcal{L}_\theta(X_t) - \log(n_T) \right) \\ &\quad - S_{\max_{\theta \in \Theta_D}} \left( \mathcal{L}_\theta(X_t) - \log(n_D) \right),\end{aligned}\quad (20)$$

which proves Equation 5.

Now we rederive  $S(X_t)$  for homogeneous visual search ( $M=L>1$ ,  $n_T=n_D=1$ ) from the local orientation log likelihoods  $\mathcal{L}_\theta(X_t^l)$  from each of the  $L$  locations. Call  $l_T \in \{1, 2, \dots, L\}$  the target location and assume uniform prior on  $l_T$ . Equation 4 is proved below:

$$\begin{aligned}&\log \frac{P(X_t|C=1)}{P(X_t|C=0)} \\ &= \log \frac{\sum_{l_T} P(X_t|l_T) P(l_T|C=1)}{P(X_t|C=0)} \\ &= \log \frac{1}{L} \sum_{l_T} \frac{P(X_t|l_T)}{P(X_t|C=0)} \\ &= \log \frac{1}{L} \sum_{l_T} \frac{P(X_t^{l_T}|\theta_T) \prod_{l \neq l_T} P(X_t^l|\theta_D)}{\prod_l P(X_t^l|\theta_D)} \\ &= \log \frac{1}{L} \sum_{l_T} \frac{P(X_t^{l_T}|\theta_T)}{P(X_t^{l_T}|\theta_D)} \\ &= S_{\max_{l_T}} \left( \mathcal{L}_{\theta_T}(X_t^{l_T}) - \mathcal{L}_{\theta_D}(X_t^{l_T}) - \log(L) \right).\end{aligned}\quad (21)$$

## Formulating common search problems using the general model

The heterogeneous visual search model is a general model for explaining a wide range of search tasks. The general model captures the variability in set size and orientation difference using CDD, which is the distribution  $P(Y^l|C^l=0)$  of stimulus orientation at a nontarget location. Below are three examples.

- Experiment 3: The distractor orientation is sampled uniformly from  $\{20^\circ, 30^\circ, 45^\circ\}$ , and all the distractors must have the same orientation. In this case a CDD is a three-dimensional vector of

$$\phi = [P(Y^l = 20^\circ | C^l = 0), P(Y^l = 30^\circ | C^l = 0), P(Y^l = 45^\circ | C^l = 0)].$$

We employ three CDDs:

$$\phi^{(1)} = [1, 0, 0]; \phi^{(2)} = [0, 1, 0]; \phi^{(3)} = [0, 0, 1];$$

with equal prior probability  $P(\phi^{(i)}) = 1/3, \forall i$ .

This setup exactly describes the probabilistic structure of Experiment 3. Since each CDD is a delta function at a single orientation, distractors at all locations will be identical.

- The distractor orientation is sampled from  $\{20^\circ, 30^\circ, 45^\circ\}$  with probability  $[0.2, 0.5, 0.3]$ .

This is the i.i.d.-distractor heterogeneous search task (Equation 9). Only one CDD is needed, and  $\phi = [0.2, 0.5, 0.3]$ .

- Experiment 2: The distractor orientation is  $30^\circ$ . The set size  $M$  is sampled uniformly from  $\{3, 6, 12\}$ . The total number of display locations is  $L = 12$ .

In this case, denote  $Y^l = \emptyset$  that a nontarget location is blank, i.e., it does not contain a stimulus bar. If there are  $M$  display items, then the probability of any nontarget location being blank is  $(L - M)/L$ . A CDD is a two-dimensional vector of

$$\phi = [P(Y^l = 20^\circ | C^l = 0), P(Y^l = \emptyset | C^l = 0)],$$

and the three different set sizes may be represented by three CDDs of equal probability:

$$\begin{aligned} \phi^{(1)} &= [3/12, 9/12], \phi^{(2)} = [6/12, 6/12], \phi^{(3)} \\ &= [1 - \epsilon, \epsilon], \end{aligned} \quad (22)$$

where  $\epsilon$  is a small number to prevent zero probability.

Note that the setup in Equation 22 only approximates the probabilistic structure of Experiment 2. This is because the blank placements are not independent of one another. In other words, for a given set size  $M$ , only  $M$  locations can contain a distractor. If we place a distractor at each location with probability  $M/L$ , we do not always observe  $M$  distractors. Instead, the actual set size follows a binomial distribution with mean  $M$ . However, this is a reasonable approximation because the human visual system can generalize to unseen set sizes effortlessly. In addition, the values of  $M$  used in our experiments are often different enough  $\{3, 6, 12\}$  that the i.i.d. model is equally effective in inferring  $M$  (Figure 4i).

## Bayesian inference for heterogeneous visual search

SPRT relies on computing  $S(X_t)$  from the orientation log likelihoods  $\mathcal{L}_\theta(X_t^l)$  from all locations  $l$ , which we show below. The target-present likelihood  $P(X_t | C = 1)$  is given by marginalizing out the target location  $l_T \in \{1, 2, \dots, L\}$ , CDD  $\phi$ , and the target and distractor orientations. Let  $C^l \in \{0, 1\}$  denote the stimulus class at location  $l$ ;  $C^l = 1$  if and only if location  $l$  contains a target. In light of the graphical model in Figure 1b,

$$\begin{aligned} P(X_t | C = 1) &= \sum_{l_T, \phi} P(X_t | l_T, \phi, C = 1) P(\phi) P(l_T | C = 1) \\ &= \sum_{l_T} P(l_T | C = 1) \sum_{\phi} P(\phi) \sum_{\vec{Y} = \{Y^1, \dots, Y^L\}} P(X_t | \vec{Y}) P(\vec{Y} | l_T, \phi, C = 1) \\ &= \sum_{l_T} P(l_T | C = 1) \sum_{\phi} P(\phi) \sum_{\vec{Y}} \prod_l (P(X_t^l | Y^l) P(Y^l | l_T, \phi, C = 1)) \\ &= \sum_{l_T} P(l_T | C = 1) \sum_{\phi} P(\phi) \prod_l \sum_{Y^l} (P(X_t^l | Y^l) P(Y^l | l_T, \phi, C = 1)) \\ &= \sum_{l_T} P(l_T | C = 1) \sum_{\phi} P(\phi) P(X_t^{l_T} | C^{l_T} = 1) \prod_{l \neq l_T} P(X_t^l | \phi, C^l = 0) \\ &= \sum_{l_T} P(l_T | C = 1) \sum_{\phi} \frac{P(X_t^{l_T} | C^{l_T} = 1)}{P(X_t^{l_T} | \phi, C^{l_T} = 0)} P(\phi) \prod_l P(X_t^l | \phi, C^l = 0), \end{aligned} \quad (23)$$

where

$$P(X_t^l | C^l = 1) = \sum_{\theta \in \Theta_T} P(X_t^l | Y^l = \theta) P(\theta | C^l = 1)$$

$$P(X_t^l | \phi, C^l = 0) = \sum_{\theta \in \Theta_D} P(X_t^l | Y^l = \theta) \phi_\theta.$$

Similarly, the target-absent likelihood is

$$P(X_t | C = 0) = \sum_{\phi} P(\phi) \prod_l P(X_t^l | \phi, C^l = 0). \quad (24)$$

Note that Equation 24 may be thought of as computing a normalization of the term  $P(\phi) \prod_l P(X_t^l | \phi, C^l = 0)$  that is used to weight the local log likelihood ratios in Equation 23. This normalized weight turns out to be the posterior of CDD:  $P(\phi | X_t)$ . Define the log posterior of CDD as

$$\begin{aligned} Q_\phi(X_t) &\triangleq \log P(\phi | X_t) \\ &= \log \frac{P(\phi) \prod_l P(X_t^l | \phi, C^l = 0)}{\sum_{\phi'} P(\phi') \prod_l P(X_t^l | \phi', C^l = 0)}. \end{aligned} \quad (25)$$

Then the log likelihood ratio is

$$\begin{aligned} &\log \frac{P(X_t | C = 1)}{P(X_t | C = 0)} \\ &= \log \sum_l P(l_T = l | C = 1) P(X_t^l | C^l = 1) \\ &\quad \times \sum_{\phi} \frac{P(\phi | X_t)}{P(X_t^l | \phi, C^l = 0)}. \end{aligned}$$

Recall that  $\text{Smax}_{i \in A} x_i = \log \sum_{i \in A} \exp(x_i)$

$$\begin{aligned} &\log \frac{P(X_t | C = 1)}{P(X_t | C = 0)} \\ &= \text{Smax}_{l=1, \dots, L} \left( \log P(l_T = l | C = 1) + \log P(X_t^l | C^l = 1) \right. \\ &\quad \left. + \text{Smax}_{\phi \in \Phi} \left( Q_\phi(X_t) - \log P(X_t^l | \phi, C^l = 0) \right) \right) \end{aligned} \quad (26)$$

assuming uniform prior on the target location  $P(l_T = l | C = 1)$  and on the target type  $P(Y^l = \theta | C^l = 1)$ :

$$\begin{aligned} &\log \frac{P(X_t | C = 1)}{P(X_t | C = 0)} \\ &= \text{Smax}_{l=1, \dots, L} \left( \text{Smax}_{\theta \in \Theta_T} \left( \mathcal{L}_\theta(X_t^l) - \log(n_T) \right) \right. \\ &\quad \left. + \text{Smax}_{\phi \in \Phi} \left( - \text{Smax}_{\theta \in \Theta_D} \left( \mathcal{L}_\theta(X_t^l) + \log \phi_\theta \right) \right. \right. \\ &\quad \left. \left. + Q_\phi(X_t) \right) - \log(L) \right), \end{aligned} \quad (27)$$

which proves Equations 10 and 11.

## Mean-field approximation to SPRT

Instead of inferring the CDD on a trial-by-trial basis, a simpler alternative is to use its average value without looking at the stimulus. For example, in the mixed set-size example with  $M \in \{3, 6, 12\}$ , SPRT estimates the value of  $M$  given  $X_t$  for each trial, whereas the simple model assumes a set size of  $\mathbb{E}(M) = 7$  for all the trials.

In detail, the simple model essentially uses the mean-field approximation on Equation 27:

$$\begin{aligned} &\log \frac{P(X_t | C = 1)}{P(X_t | C = 0)} \\ &\approx \text{Smax}_{l=1, \dots, L} \left( \text{Smax}_{\theta \in \Theta_T} \left( \mathcal{L}_\theta(X_t^l) \right) \right. \\ &\quad \left. - \text{Smax}_{\theta \in \Theta_D} \left( \mathcal{L}_\theta(X_t^l) + \log \bar{\phi}_\theta \right) \right) - \log(n_T L), \end{aligned} \quad (28)$$

where  $\bar{\phi}_\theta = \sum_{\phi \in \Phi} \phi_\theta P(\phi)$  is the mean CDD with respect to the prior distribution. The prediction of the simple model on a mixed set-size search problem is shown in Figure 6b.

## Search with correlated target and distractor orientations

SPRT for heterogeneous visual search (Equation 27) assumes that the properties of the scene—namely the set size and the scene complexity—affect only the distractor orientation distribution. In this section we relax this assumption and let  $\phi$  encode both the target and distractor orientation distribution:  $\phi = \{\phi^{(T)}, \phi^{(D)}\}$ , where  $\phi_\theta^{(T)} = P(Y^l = \theta | C^l = 1)$  and  $\phi_\theta^{(D)} = P(Y^l = \theta | C^l = 0)$ . The log likelihood of target present in Equation 23 now becomes

$$\begin{aligned} &P(X_t | C = 1) \\ &= \sum_{l_T} P(l_T | C = 1) \sum_{\phi} \frac{P(X_t^{l_T} | \phi^{(T)}, C^{l_T} = 1)}{P(X_t^{l_T} | \phi^{(D)}, C^{l_T} = 0)} \\ &\quad \times P(\phi) \prod_l P(X_t^l | \phi^{(D)}, C^l = 0). \end{aligned}$$

The log likelihood ratio of target present versus target absent is



$$\begin{aligned} \log \frac{P(X_t|C=1)}{P(X_t|C=0)} &= \log \sum_l P(l_T = l|C=1) \sum_{\phi} \frac{P(X_t^l|\phi^{(T)}, C^l=1)}{P(X_t^l|\phi^{(D)}, C^l=0)} P(\phi|X_t) \\ &= \mathcal{S}\max_{l=1,\dots,L} (\mathcal{S}\max_{\phi \in \Phi} (\mathcal{S}\max_{\theta \in \Theta_T} (\mathcal{L}_{\theta}(X_t^l) + \log \phi_{\theta}^{(T)}) - \mathcal{S}\max_{\theta \in \Theta_D} (\mathcal{L}_{\theta}(X_t^l) + \log \phi_{\theta}^{(D)}) + Q_{\phi}(X_t))) - \log(L). \end{aligned} \quad (29)$$

This formulation encompasses the formulation in Equation 27 where the target and the distractor orientations are distributed independently with respect to each other. To see this, assume  $\phi^{(D)}$  and  $\phi^{(T)}$  vary independently. Then,

$$\begin{aligned} P(X_t|C=1) &= \sum_{l_T} P(l_T|C=1) \sum_{\phi^{(T)}, \phi^{(D)}} \frac{P(X_t^{l_T}|\phi^{(T)}, C^{l_T}=1)}{P(X_t^{l_T}|\phi^{(D)}, C^{l_T}=0)} P(\phi^{(T)}) P(\phi^{(D)}) \prod_l P(X_t^l|\phi^{(D)}, C^l=0) \\ &= \sum_{l_T} P(l_T|C=1) \sum_{\phi^{(D)}} \frac{\sum_{\phi^{(T)}} P(\phi^{(T)}) P(X_t^{l_T}|\phi^{(T)}, C^{l_T}=1)}{P(X_t^{l_T}|\phi^{(D)}, C^{l_T}=0)} P(\phi^{(D)}) \prod_l P(X_t^l|\phi^{(D)}, C^l=0) \\ &= \sum_{l_T} P(l_T|C=1) \sum_{\phi^{(D)}} \frac{P(X_t^{l_T}|\bar{\phi}^{(T)}, C^{l_T}=1)}{P(X_t^{l_T}|\phi^{(D)}, C^{l_T}=0)} P(\phi^{(D)}) \prod_l P(X_t^l|\phi^{(D)}, C^l=0), \end{aligned} \quad (30)$$

where  $\bar{\phi}^{(T)} = \sum_{\phi^{(T)}} \phi^{(T)} P(\phi^{(T)})$  is the expected value of  $\phi^{(T)}$ . Equation 30 is equivalent to Equation 27 with a different prior ( $\bar{\phi}^{(T)}$ ) on target orientation.

## Search with multiple targets

One may derive the SPRT under a different hypothesis: that a target may appear at each location with i.i.d. probability  $p_T$ ; for example,  $p_T = 1 - 0.5^{1/M}$  will produce target-absent scenes with probability 0.5. For the simple case of homogeneous visual search—known set size  $M$ , unique target orientation  $\Theta_T = \{\theta_T\}$ , and distractor orientation  $\Theta_D = \{\theta_D\}$ —we first derive the likelihood of observations with and without the class label:

$$\begin{aligned} P(X_t|C=0) &= \prod_{l=1}^M P(X_t^l|C^l=0) \\ P(X_t) &= \prod_{l=1}^M P(X_t^l) = \prod_{l=1}^M (P(X_t^l|C^l=0)P(C^l=0) \\ &\quad + P(X_t^l|C^l=1)P(C^l=1)). \end{aligned}$$

Recall that  $S(X_t^l) = \log(P(X_t^l|C^l=1)/P(X_t^l|C^l=0))$  is the log likelihood ratio at location  $l$ , and let  $R_0 \triangleq \log\left(\frac{p_T}{1-p_T}\right)$  be the log prior ratio at a single location. The log posterior of target absent is

$$\begin{aligned} \log P(C=0|X_t) &= \log \frac{P(X_t|C=0)P(C=0)}{P(X_t)} \\ &= \log \prod_{l=1}^M \frac{P(X_t^l|C^l=0)P(C^l=0)}{P(X_t^l|C^l=0)P(C^l=0) + P(X_t^l|C^l=1)P(C^l=1)} \\ &= \log \prod_{l=1}^M \frac{1}{1 + \exp(S(X_t^l) + R_0)} \\ &= \sum_{l=1}^M -\log(1 + \exp(S(X_t^l) + R_0)). \end{aligned}$$

The log posterior ratio of target present versus target absent is then

$$\begin{aligned} \log \frac{P(C=1|X_t)}{P(C=0|X_t)} &= \log(\exp(-\log P(C=0|X_t)) - 1) \\ &= \text{bow}_{l=1,\dots,M}(S(X_t^l) + R_0), \end{aligned} \quad (31)$$

where

$$\text{bow}_{i \in A}(x) \triangleq \log\left(\exp\left(\sum_{i \in A} \log(\exp(x_i) + 1)\right) - 1\right)$$

is the “bow” function (our name), related to the Fermi-Dirac equation of statistical physics (Reif, 1965). This function behaves like a softmax for the negative inputs and a sum for the positive inputs. The bow function can be approximated by the “sum-max” function that sums the positive components of the inputs and maximizes

over the negative components:

$$\text{sum-max}(x) \triangleq \sum_{i \in A} (x_i)_+ + \max_{i \in A} ((x_i)_-),$$

where  $(x)_+$  and  $(x)_-$  denote the positive and negative parts of  $x$ , respectively.

## Gain control

Gain control over a population of neurons, where each neuron  $i$  has membrane potential  $U_i$ , is given by

$$\tilde{U}_i = U_i - g(U), \quad (32)$$

where  $g(\cdot)$  is a function computing the gain of the population. Gain control serves two purposes: one to control the range of a neuron's membrane potential within its physiological limits (as needed in Equation 12), and the other to provide normalization to allow a probabilistic interpretation of the population code (as needed in Equation 25). One gain that serves both purposes is the softmax function:  $g(U) = \mathcal{S}\text{max}(U_i)$ , which is what we use for computing the log posterior of CDD  $Q_\phi(X_t)$  (Equation 25). While the popular form of gain control is divisive normalization (Carandini & Heeger, 2011), we use subtractive normalization (Doiron, Longtin, Berman, & Maler, 2001) because it comes out naturally from the SPRT computation involving log likelihoods.

## Unary encoding with spikes

To communicate a neuron's continuous, time-dependent membrane potential  $U(t)$  to a distant neuron, we hypothesize that nature makes use of action potentials (unary encoding in engineering). The sender neuron maintains two thresholds  $\tau_0^s < 0 < \tau_1^s$ . Whenever  $U(t)$  shoots above the positive threshold  $\tau_1^s$ , a spike is generated and the membrane discharges by an amount equal to the threshold; that is,  $U(t+1) = U(t) - \tau_1^s$ . If a discharged membrane potential  $U(t')$  is still bigger than the threshold, then another spike is generated after a refractory period  $t_{ref}$ :  $U(t'+t_{ref}) = U(t') - \tau_1^s$ ; this process is repeated, generating a burst of  $k$  spikes, until  $U(t' + kt_{ref}) < \tau_1^s$ . The spikes travel to the receiver neuron through an excitatory synapse whose strength is equal to the threshold, thus allowing the receiver to compute  $U(t)$  with only minimal delay. Similar mechanisms can be implemented for the case where  $U(t)$  drops below  $\tau_0^s$  (see Supplementary Figure S2 and Figure 4g through i).

**Keywords:** visual search, ideal observer, speed–accuracy tradeoff, spiking neural networks, gain control

## Acknowledgments

The authors acknowledge comments from and discussions with Jeremy M. Wolfe, Jeffrey D. Schall, and Ueli Rutishauser. This work was funded by ONR N00014-10-1-0933 and Gordon and Betty Moore Foundation.

Commercial relationships: none.

Corresponding author: Pietro Perona.

Email: perona@caltech.edu.

Address: Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA.

## References

- Avraham, T., Yeshurun, Y., & Lindenbaum, M. (2008). Predicting visual search performance by quantifying stimuli similarities. *Journal of Vision*, 8(4):9, 1–22, doi:10.1167/8.4.9. [PubMed] [Article]
- Beck, J., Ma, W., Kiani, R., Hanks, T., Churchland, A., Roitman, J., . . . Pouget, A. (2008). Probabilistic population codes for Bayesian decision making. *Neuron*, 60(6), 1142–1152.
- Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1), 1–127.
- Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4), 700–765.
- Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, 57(3), 153–178.
- Busmeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, 32(2), 91–134.
- Busmeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3), 432–459.
- Cameron, E. L., Tai, J. C., Eckstein, M. P., & Carrasco, M. (2004). Signal detection theory applied to three visual search tasks—Identification, yes/no detection and localization. *Spatial Vision*, 17(4), 295–326.
- Carandini, M., & Heeger, D. J. (2011). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1), 51–62.

- Carandini, M., Heeger, D. J., & Movshon, J. A. (1999). Linearity and gain control in v1 simple cells. In E. G. Jones & A. Peters (Eds.), *Models of cortical circuits* (pp. 401–443). New York, NY: Springer.
- Carrasco, M., & Yeshurun, Y. (1998). The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2), 673–692.
- Cassey, P., Heathcote, A., & Brown, S. D. (2014). Brain and behavior in decision-making. *PLoS Computational Biology*, 10(7), e1003700.
- Chen, B., Navalpakkam, V., & Perona, P. (2011). Predicting response time and error rates in visual search. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 2699–2707). Granada, Spain: NIPS Foundation.
- Chen, B., & Perona, P. (2014). Towards an optimal decision strategy of visual search. preprint arXiv: 1411.1190. Accessed 1 Nov 2014.
- Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2), 199–213.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.
- Davis, E. T., & Graham, N. (1981). Spatial frequency uncertainty effects in the detection of sinusoidal gratings. *Vision Research*, 21(5), 705–712.
- Dayan, P., & Abbott, L. (2003). Theoretical neuroscience: Computational and mathematical modeling of neural systems. *Journal of Cognitive Neuroscience*, 15(1), 154–155.
- Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1), 193–222.
- DeValois, R. L., & DeValois, B. K. K. (1980). Spatial vision. *Annual Review of Psychology*, 31, 309–341.
- Doiron, B., Longtin, A., Berman, N., & Maler, L. (2001). Subtractive and divisive inhibition: Effect of voltage-dependent inhibitory conductances and noise. *Neural Computation*, 13(1), 227–248.
- Doshier, B. A., Han, S., & Lu, Z.-L. (2004). Parallel processing in visual search asymmetry. *Journal of Experimental Psychology: Human Perception and Performance*, 30(1), 3–27.
- Doshier, B. A., Han, S., & Lu, Z.-L. (2010). Information-limited parallel processing in difficult heterogeneous covert visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 36(5), 1128–1144.
- Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *The Journal of Neuroscience*, 32(11), 3612–3628.
- Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3), 433–458.
- Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision*, 11(5):14, 1–36, doi:10.1167/11.5.14. [PubMed] [Article]
- Eckstein, M. P., & Abbey, C. K. (2001). Model observers for signal-known-statistically tasks (SKS). *Proceedings of SPIE*, 4324, 91–102, doi:10.1117/12.431177.
- Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1), 1–47.
- Gabbiani, F., & Koch, C. (1996). Coding of time-varying signals in spike trains of integrate-and-fire neurons with random threshold. *Neural Computation*, 8(1), 44–66.
- Geisler, W. S. (1989). Sequential ideal-observer analysis of visual discriminations. *Psychological Review*, 96(2), 267–314.
- Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research*, 51(7), 771–781.
- Goris, R. L., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, 17(6), 858–865.
- Graf, A. B., Kohn, A., Jazayeri, M., & Movshon, J. A. (2011). Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature Neuroscience*, 14(2), 239–245.
- Gray, C. M., & McCormick, D. A. (1996). Chattering cells: Superficial pyramidal neurons contributing to the generation of synchronous oscillations in the visual cortex. *Science*, 274(5284), 109–113.
- Green, D., & Swets, J. (1966). *Signal detection theory and psychophysics*. Los Altos, CA: Peninsula.
- Heitz, R. P., & Schall, J. D. (2012). Neural mechanisms of speed-accuracy tradeoff. *Neuron*, 76(3), 616–628.
- Hodson, J., & Humphreys, G. W. (2001). Driving attention with the top down: The relative contribution of target templates to the linear separability effect in the size dimension. *Perception & Psychophysics*, 63(5), 918–926.
- Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in

- the cat's visual cortex. *The Journal of Physiology*, 160(1), 106–154.
- Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience*, 9(5), 690–696.
- Koch, C., & Ullman, S. (1987). Shifts in selective visual attention: Towards the underlying neural circuitry. In L. M. Vaina (Ed.), *Matters of intelligence* (pp. 115–141). New York, NY: Springer.
- Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing* (pp. 1106–1114). Lake Tahoe: NIPS Foundation.
- Lo, C.-C., & Wang, X.-J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience*, 9(7), 956–963.
- Ma, W. J., Navalpakkam, V., Beck, J. M., Van Den Berg, R., & Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. *Nature Neuroscience*, 14(6), 783–790.
- Mazurek, M. E., Roitman, J. D., Ditterich, J., & Shadlen, M. N. (2003). A role for neural integrators in perceptual decision making. *Cerebral Cortex*, 13(11), 1257–1269.
- Moore, T., & Fallah, M. (2004). Microstimulation of the frontal eye field and its effects on covert spatial attention. *Journal of Neurophysiology*, 91(1), 152–162.
- Nagy, A. L., Neriani, K. E., & Young, T. L. (2005). Effects of target and distractor heterogeneity on search for a color target. *Vision Research*, 45(14), 1885–1899.
- Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434(7031), 387–391.
- Navalpakkam, V., Koch, C., Rangel, A., & Perona, P. (2010). Optimal reward harvesting in complex perceptual environments. *Proceedings of the National Academy of Sciences, USA*, 107(11), 5232–5237.
- Oster, M., Douglas, R., & Liu, S.-C. (2009). Computation with spikes in a winner-take-all network. *Neural Computation*, 21(9), 2437–2465.
- Palmer, E., Horowitz, T., Torralba, A., & Wolfe, J. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, 37(1), 58–71.
- Palmer, J. (1994). Set-size effects in visual search: The effect of attention is independent of the stimulus for simple tasks. *Vision Research*, 34(13), 1703–1721.
- Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, 5(5):1, 376–404, doi:10.1167/5.5.1. [PubMed] [Article]
- Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research*, 40(10), 1227–1268.
- Perona, P. (2014). Quantized response times are a signature of a neuronal bottleneck in decision. *Frontiers in Computational Neuroscience*, 8(42), 1–10.
- Pomplun, M., Garaas, T. W., & Carrasco, M. (2013). The effects of task difficulty on visual search strategy in virtual 3D displays. *Journal of Vision*, 13(3):24, 1–22, doi:10.1167/13.3.24. [PubMed] [Article]
- Posner, M., Cohen, Y., & Rafal, R. (1982). Neural systems control of spatial orienting. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 298(1089), 187–198.
- Purcell, B. A., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2012). From salience to saccades: Multiple-alternative gated stochastic accumulator model of visual search. *The Journal of Neuroscience*, 32(10), 3433–3446.
- Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses. *Psychological Review*, 92(2), 212–225.
- Reif, F. (1965). *Fundamentals of statistical and thermal physics* (Vol. 1). New York, NY: McGraw-Hill.
- Rosenholtz, R. (2001). Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal-detection theory models of search. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4), 985–999.
- Rutishauser, U., & Koch, C. (2007). Probabilistic modeling of eye movement data during conjunction search via feature-based attention. *Journal of Vision*, 7(6):5, 1–20, doi:10.1167/7.6.5. [PubMed] [Article]
- Sanger, T. D. (1996). Probability density estimation for the interpretation of neural population codes. *Journal of Neurophysiology*, 76(4), 2790–2793.
- Schwarz, W. (2001). The ex-wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments, & Computers*, 33(4), 457–469.
- Seung, H. S. (2009). Reading the book of memory:



- Sparse sampling versus dense mapping of connectomes. *Neuron*, 62(1), 17–29.
- Shadlen, M., & Newsome, W. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *Journal of Neurophysiology*, 86(4), 1916–1936.
- Shimozaki, S. S., Schoonveld, W. A., & Eckstein, M. P. (2012). A unified bayesian observer analysis for set size and cueing effects on perceptual decisions and saccades. *Journal of Vision*, 12(6):27, 1–26, doi:10.1167/12.6.27. [PubMed] [Article]
- Simoncelli, E. P., Paninski, L., Pillow, J., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. *The Cognitive Neurosciences*, 3, 327–338.
- Sperling, G., & Melchner, M. J. (1978). The attention operating characteristic: Examples from visual search. *Science*, 202(4365), 315–318.
- Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, 25(3), 251–260.
- Treisman, A., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3), 550–592.
- Van Essen, D. C., Newsome, W. T., & Maunsell, J. H. (1984). The visual field representation in striate cortex of the macaque monkey: Asymmetries, anisotropies, and individual variability. *Vision Research*, 24(5), 429–448.
- Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron*, 31(4), 523–535.
- Verghese, P., & Nakayama, K. (1994). Stimulus discriminability in visual search. *Vision Research*, 34(18), 2453–2467.
- Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, 287(5456), 1273–1276.
- Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2), 117–186.
- Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19(3), 326–339.
- Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5), 955–968.
- Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99–119). New York: Oxford University Press.
- Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6), 495–501.
- Wolfe, J. M., Horowitz, T. S., & Kenner, N. M. (2005). Rare items often missed in visual searches. *Nature*, 435(7041), 439–440.
- Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. *Vision Research*, 50(14), 1304–1311.
- Wolfe, J. M., & Van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology*, 20(2), 121–124.
- Wong, K.-F., Huk, A. C., Shadlen, M. N., & Wang, X.-J. (2007). Neural circuit dynamics underlying accumulation of time-varying evidence during perceptual decision making. *Frontiers in Computational Neuroscience*, 1, 6.
- Woodman, G. F., Kang, M.-S., Thompson, K., & Schall, J. D. (2008). The effect of visual search efficiency on response preparation: Neurophysiological evidence for discrete flow. *Psychological Science*, 19(2), 128–136.